The Dissertation Committee for Dongyang Wang
certifies that this is the approved version of the following dissertation:

Coordinating Healthcare Networks

Committee:

Douglas Morrice, Supervisor

Kumar Muthuraman, Co-Supervisor

Edward Anderson

Jonathan Bard

Luci Leykum

Coordinating Healthcare Networks

by

Dongyang Wang, B.E.; M.S.IROM

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

December 2016

To my parents

Acknowledgments

I thank my supervisors, Prof.Douglas Morrice and Prof.Kumar Muthuraman for their invaluable training and support throughout my PhD study. They are not only my advisors in academics, but also my mentors in life.

I had the pleasure of collaborating with all my committee members at different stages of my PhD training and I thank them, Prof.Jonathan Bard, Prof.Edward Anderson and Dr.Luci Leykum, for sharing their wisdom with me. I thank the late Prof.Reuben McDaniel, for his trust and support on my first healthcare project.

I thank Zicheng Hu for being my audience of research ideas for five years.

Lastly, I thank my eight furry friends in the neighborhood for their sup-purr.

Coordinating Healthcare Networks

Publication No. _____

Dongyang Wang, Ph.D.
The University of Texas at Austin, 2016

Supervisors: Douglas Morrice
Kumar Muthuraman

Current healthcare reforms advocate significantly to improve the coordination of services around a patient-centric model, with an overarching goal to maximize patient outcomes with lower cost, i.e. a value-based care. With most patient care delivered through outpatient services, the need to coordinate different services and their patient appointment scheduling decisions becomes central to successful reform. Currently, outpatient services are particularly fragmented with minimal coordination among different providers, and the coordination is left to the patient. This approach causes compromised patient health outcomes, an increase in missed appointments and unacceptable access delays. Therefore, the potential impact of coordinating outpatient services is great, in terms of improving patient outcomes and satisfaction, optimizing providers' utilization and reducing operational costs.

In the first study, we investigate how to coordinate the delivery of care in the preoperative process for surgical outpatient. Based on the concept of the Perioperative Surgical Home proposed by the American Society of Anesthesiologists, we

develop a Patient-Centered Surgical Home (PCSH) model. Using statistical analysis and simulation, we demonstrate how this can be implemented and reveal the potential benefits on cooperation of the referring clinics and integrating patient information early in the preoperative process.

The second study proposes a multi-station network model that sequentially schedules patient appointments in a network of stations with stochastic service times, no-show possibilities, and overbooking. We propose a myopic coordinated policy and present evidence that the policy yields a solution that is close to optimal and is computationally feasible. However, the solution is not simple enough for practical implementation. Hence, we explore a sequence of approximations and find one that offers a tremendous computational advantage. We also provide several managerial insights and discuss how network structures affect complexity.

In the third study, we focuses on the cost perspective of coordination. We formulate a multi-server, multi-clinic model that represents the current practice at the PCSH and develop a coordinated scheduling method that dynamically balances the utilizations of all services as patients are sequentially scheduled in the PCSH. We compare our proposed policy against other policies found in the practice and the results shed light on the risk of improper coordination in our increasingly interdependent healthcare system.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Annual spending on health in the United States is projected to grow 5.8% each year between 2014 and 2024. This growth rate is 1.1 times faster than the average GDP growth rate. It was estimated to hit $3.2 trillion dollars in 2015, which is already about 18% of the GDP (Bureau of Economic Analysis (2016); Centers for Medicare and Medicaid Services (2015)). Such spending is clearly unsustainable. Given the country's aging population, the outlook, needless to say, is troubling. Managing this increasing demand, when accompanied by tightening budget and resource scarcity, depends primarily on our ability to improve the efficiency and effectiveness of our care delivery. Recent healthcare reforms recognize this need and focus on moving toward patient-centered care to improve both the efficiency and effectiveness (Centers for Medicare and Medicaid Services (2009)) of the delivery of care. In this dissertation, we focus on the outpatient surgical care, which is particularly fragmented in the current healthcare system. The preoperative process often involves multiple providers (e.g., primary care physicians, surgeons, anesthesiologists, and nurses) depending on patients' needs, but there is minimal coordination among different providers. As a consequence, patients have to plan for and coordinate their own medical trips. Therefore, the potential impact of coordinating outpatient services is great in terms improving patient outcomes and satisfaction,

optimizing providers' utilization and reducing operational costs in outpatient surgical care. As pointed out by Porter (2009), the overarching strategy of our healthcare reform should focus on maximizing patient outcomes at the lowest cost, i.e. a value-based care, which is defined as health outcome over cost spending.

In the first study, we develop a Patient-Centered Surgical Home (PCSH) to improve the health outcome of surgical outpatients. The PCSH is based on the concept of the Perioperative Surgical Home proposed by the American Society of Anesthesiologists. A key feature of the PCSH is to have an anesthesiology preoperative assessment clinic (APC) serve as system coordinator and information integrator. Based on a study of outpatient surgery at the University of Texas Health Science Center at San Antonio and its primary teaching hospital using statistical analysis and simulation, we demonstrate how this can be accomplished. Our study reveals: i) bottlenecks in APC and its patient assessment capacity; ii) significant sources of variability, particularly in patient arrivals and patient assessment times; and iii) opportunities for process improvement, especially with regard to information deficiencies (Lahiri and Seidmann (2012)) and patient screening. Our analysis shows that with the proper screening tool and modifications to the way triage is handled, it is possible to increase the number of patients that the APC sees each day with a modest increase in resources. Much of the potential benefits rest on the cooperation of the referring clinics as well as closing the gap between the current level of patient information and what is needed for optimizing medical decisions. Since APC-like clinics are common in practice, our findings have great potential for widespread implementation of similar PCSH models with commensurate benefits.

The second study proposes a methodology to coordinate multiple clinics in a healthcare network via patient appointments scheduling. The goal is to improve patient access to care and reduce operational costs due to system uncertainties, via better coordination. An advantage of coordinated scheduling is the opportunity to anticipate and accommodate referrals before the patients' arrival. With this type of advanced planning, it is possible to schedule multiple services on a single patient visit, improving access to care and patient satisfaction. It also has the potential to reduce patient no-shows associated with referral appointments, mitigating uncertainty and thus reducing operational inefficiencies and costs in healthcare. Coordinated appointment scheduling is very limited in literature and to our knowledge, implementable models do not exist. As pointed out by Berg and Denton (2012), managing healthcare as a multi-station interconnected network is an open and important problem in the OR/Healthcare Management research. In this chapter, we fill the technological gap in accomplishing coordinated scheduling.

We formulate a stochastic network model that captures the complexity of patient no-shows, sequential scheduling necessity, service time uncertainty, and stochastic patient flows within and between services. A centralized scheduler uses patient information and preferences to make sequential appointment decisions that maximize a network objective. The objective is to balance the benefit of serving patients against the costs involved in patient waiting time and staff overtime. We propose a coordinated myopic policy and we show that for a range of parameters, our myopic approach yields solutions that are within 1% of an unachievable super-optimal solution. In addition, for practical implementation, we create a number of approx-

imation schemes searching for ones that yield large computational advantages at very low approximation costs, and manage to find one such very beneficial approximation.

The third study focuses on the cost perspective of the coordinated scheduling and is motivated by the challenge facing the PCSH on how to coordinate APC and IMC. We develop a multi-server, multi-clinic model that represent the current operations at the PCSH. We propose a joint scheduling approach to sequentially allocate patient requests to the best appointment slots that maximizes the objective of the PCSH. Because APC patients are of higher priority and their referral services are guaranteed on the same day, we develop a simple heuristic to ensure timely access of care for APC, as well as IMC patients. By comparing our proposed policy against other policies used or considered by the PCSH, we address the pressing questions posed by clinical practitioners: (1) How much operational inefficiencies can be reduced by coordination? (2) What is the opportunity cost if clinics continue with their current scheduling policies, which do not fully support coordination? The results in our numerical study shed light on the risk in our increasingly interdependent healthcare system, if coordination is not properly implemented.

Chapter 2

A Patient-Centered Surgical Home to Improve Outpatient Surgical Processes of Care and Outcomes*

2.1. Introduction

Improving the safety and efficiency of healthcare delivery is critical as healthcare reimbursement moves to outcomes-based standards that rely on integration of care across providers. Surgical care can be particularly fragmented. Several types of physicians are involved in the care of these patients: surgeons, anesthesiologists, primary care physicians, specialists, and increasingly, hospitalists. During the surgical episode itself, these physicians must coordinate with nursing and other services such as physical therapy. Thus, the original system is vulnerable to lack of coordination, fragmentation, inefficiencies, patient information deficiencies, system congestion, and delays and cancellations. These issues can in turn lead to inappropriate utilization of services or suboptimal clinical outcomes.

In recognition of the need for better coordination of care, the American Society of Anesthesiologists (ASA) has conceptualized the "Perioperative Surgical Home (PSH)" (ASA (2013)). This proposal states that under a PSH, care is integrated and coordinated among specialties, with anesthesiologists serving as system coordinators and information integrators, in collaboration with general internists/hospital-

* Morrice, D. J., D. Wang, J. F. Bard, L. K. Leykum, S. Noorily, and P. Veerapaneni (2014). A patient-centered surgical home to improve outpatient surgical processes of care and outcomes. IIE Transactions on Healthcare Systems Engineering 4, 119–134. Dongyang Wang is the main author of this paper, who contributed to the simulation model, data analysis and the results.

ists and other physicians. We have chosen to refer to the concept of a PSH as a "Patient-Centered Surgical Home (PCSH)") in order to emphasize the centrality of the patient in this system of care. While this approach has conceptual appeal (Vetter et al. (2013a); Vetter et al. (2013b)), it has not yet been widely implemented or studied because it represents a significantly different way of delivering perioperative care.

Changing the model of care delivery for surgical patients has taken on increasing urgency. The frequency and importance of outpatient (or ambulatory) surgical care has grown significantly in recent years (Cullen et al. (2009)) due to benefits from lower complication and infection rates, patient convenience, and lower costs (Berg and Denton (2012)). In this chapter, we develop a PCSH model for outpatient surgery. We redesign the preoperative processes of care to create a unified pathway that is treated like a single episode with better provider coordination. From an Operations Management perspective, the PCSH represents a systems approach to delivering outpatient surgical care.

Our motivating case study centers on outpatient surgery at the University of Texas Health Science Center at San Antonio School of Medicine faculty practice, UT Medicine (UTM), and University Hospital, the acute care facility for University Health System (UHS), UTM's primary teaching affiliate. We focus on this case study throughout the chapter because it exemplifies issues that are prevalent in the literature and in practice. Patients enter UHS outpatient surgery by visiting one of the eighteen UTM or UHS surgical clinics. Those requiring surgery are scheduled for outpatient surgery at UHS. In the original system, about 40% of these patients are

referred to the Anesthesia Preoperative Clinic (APC) for an assessment prior to their day of surgery to make sure that they are medically ready for the procedure. Presumably those with more complicated medical conditions are referred to APC; the remaining patients undergo an assessment on the day of surgery. We conducted a system-level analysis of outpatient surgery that revealed two important issues. First, we found a lack of coordination between the surgery clinics and APC. Second, this lack of coordination led to operational and potentially clinical impacts at the time of surgery. APC has the potential to mitigate these issues by assessing all patients prior to the day of surgery. Furthermore, it could take the lead and serve as system coordinator and information integrator for the PCSH model, so the system could function in a more unified fashion as suggested in the ASA (ASA (2011)) proposal, with a modest increase in resource requirements. The value of a systems coordinator (or "integrator") has been demonstrated in other Operations Management literature by Parker and Anderson Jr. (2002).

Our work builds on numerous studies that show the benefits of APC-like clinics on various activities in this system (Newman et al. (2013)). However, the endeavor is not without challenges. In order for this approach to work, one must ensure that APC sees the right patients with the right information. By conducting an extensive study involving process analysis, statistical analysis, and simulation, we demonstrate how this can be accomplished. More specifically, the study reveals: i) bottlenecks in APC and its patient assessment capacity; ii) significant sources of variability, particularly in patient arrivals and patient assessment times; and iii) opportunities for process improvement, especially with regard to information deficiencies

(Lahiri and Seidmann (2012)) and patient screening. Patient information deficiency is one of the key factors in our study. Therefore, we quantify its impact on providers' time in the APC along with two other potentially important factors: patient complexity (determined by the number of comorbidities) and surgical complexity. Then we demonstrate the potential benefits that result from mitigating information deficiency through a new patient screening process coordinated and managed by APC. The new process requires APC to assess all surgical outpatients, with a triage in which more complex patients are seen in the clinic and the rest are interviewed over the telephone. By improving the patient screening process and reducing information deficiency, we show that APC's scarce resources can be utilized more effectively. In fact, with only a modest increase in resources, it can provide appropriate assessments for all patients. For APC, this represents an average increase in demand for those patients needing to be seen in the clinic of about 33% and an average increase in overall demand of about 250%.

Given that APC is central to our PCSH model, most of the research questions are related to the functioning of APC and its interactions with other parts of the system. Specifically, we seek to address the following questions:

1. What capabilities would APC need to serve as the PCSH system coordinator and information integrator? If all surgical patients were required to undergo an assessment prior to the day of surgery, what additional resources would be required to handle the increased load?

2. What is the impact of patient complexity, surgical complexity, and information

deficiency on the functioning of APC? What is the relationship between these factors and the additional time the staff must spend on each patient?

3. If an effective and acceptable screening tool can be developed resulting in a correct patient triage and reduced information deficiency, what impact would that have on the functioning of APC?

4. Can APC's performance benefit from alternative patient scheduling rules, enforcing a patient arrival window around scheduled appointment times, and staggered start times for staff?

5. What are the cost implications of moving from a standard APC to a PCSH?

The potential is great. In addition to improvements in patient care, the estimated cost savings to UHS in better operating room (OR) utilization and decreased unnecessary patient testing is over one million dollars per year. Since APC-like clinics are common in practice, our findings have great potential for widespread implementation of similar PCSH models with commensurate benefits.

The rest of the chapter is organized in the following manner. Section 3.2 provides a literature review. Section 3.3 discusses the need for a PCSH system with descriptions of the original and new systems. In Section 3.4, we describe the APC simulation model and validate this model on the original system. Section 3.5 contains a simulation analysis of an APC-coordinated PCSH model in which we evaluate the benefits of mitigating patient information deficiencies, proper patient triage, different patient scheduling rules, and a few other factors. In Section 3.6, we provide

a discussion on important insights from this work. We conclude with a summary and discussion of future work in Section 3.7.

2.2. Literature Review

The PCSH model (ASA (2013)) has its roots in the Patient Centered Medical Home (PCMH) model, a concept that originated in primary care. The main PCMH features include "comprehensiveness, integration/coordination, relationships involving sustained partnership, and new ways of organizing practice" (Stange et al. (2010)). While the PCSH incorporates these features (i.e., the model would be patient-centered and provide continuity of care), it is more limited in scope because outpatient surgery is a more focused and time-limited episode of care than longitudinal primary care. On the other hand, a surgical episode of care may be higher risk. The perioperative period has great potential for complications that can have devastating clinical impact. Surgical care is technology and resource intensive, and complications can be costly. In the current reimbursement climate, surgical care provides disproportionate revenue when compared with medical care. In the case of UTM, it accounts for 49% of revenue. In more outcomes-based or population-health-based payment models, ensuring efficiency and safety of high cost surgical care will be important.

Much has been published on APC-like clinics, which are sometimes referred to as Pre-anesthesia (or Preoperative) Assessment (or Evaluation) Clinics (see, e.g., Edward et al. (2008); Zonderland et al. (2009)), and their value has been firmly established (Newman et al. 2013). Specifically, such clinics have been shown to improve

operating room efficiency (Correll et al. (2006)), reduce unnecessary tests and consultations (Tsen et al. (2002)), reduce operating room cancellations and delays and length of hospital stay (van Klei et al. (2002); Ferschl et al. (2005)), optimize postoperative outcomes (Halaszynski et al. (2004)), and increase patient satisfaction (Hepner et al. (2004)). In each of these works, the focus is usually on one aspect of the system. While most of our analysis focuses on improving APC, it is with the primary objective of elevating APC's role as systems coordinator in the PCSH.

Zonderland et al. (2009) apply queuing analysis to improve the performance of an APC. Their study differs from ours in a number of ways. While their work provides some guidance on process improvement, they make strong assumptions (e.g., the system reaches steady state, the stages in the process are separable, and the patients arrive on time) for the particular queuing model they use in their analysis. Consequently, their results are conservative, generally over-estimating clinic waiting times. Further, they do not take a systems approach congruent with a PCSH model, but rather assume that information gathering and tests for patients are completely exogenous to their mission (patients are sent off to collect this information, get tests, etc.). Lastly, they do not link information deficiency to the clinic times. Hence, after they develop the improved clinic configuration, they use the same assessment times as in the original clinic.

Vetter et al. (2013a) provide an overview of the surgical home at the University of Alabama at Birmingham. Their work differs from ours because it is conceptual without significant analysis of the system. Additionally, their model is not integrated and coordinated among specialties as the model we implement at our institu-

tion. For example, Vetter et al. (2013a) focus more on the importance of staffing the surgical home with anesthesiologists and less on the collaboration between anesthesiology, the surgical clinics, and other areas of medicine. Vetter et al. (2013b) provide a conceptual framework and a set of analysis methodologies for evaluating the effectiveness of a PSH.

Lahiri and Seidmann (2012) demonstrate the importance of timely information collection based on an empirical and queuing analysis of a network of radiology clinics. They argue that collecting information upstream in a process well in advance of its usage avoids information deficiencies (or "information hang-overs") later in the process, when it often causes prolonged delays and may lead to negative health consequences for patients. Gibby and Schwab (1998) examine the prevalence of missing internal and external information in an outpatient APC and quantify its impact on provider consult times. As Dexter (1999) points out, rectifying this type of information deficiency has the potential to decrease patient waiting times. Using simulation, Edward et al. (2008) show that the time patients spend waiting before consultation at an APC can be reduced by making scheduled consultation times dependent on the health of the patient. Our work builds on these studies by specifically quantifying the impact of information deficiency on providers' time (both direct consult time and follow-up) in the APC and comparing its impact to other factors such as the patient complexity and the complexity of the surgical procedure.

2.3. The Need For a PCSH

The PCSH is motivated by improved patient care, better utilization of the operation rooms (OR), and more appropriate tests for patients. For example, the UHS administration estimates that lost OR time due to incomplete anesthesia preoperative assessments costs UHS approximately $13,500 per month in lost revenue. Additionally, in a recent UTM/UHS study on tests ordered for patients sent to APC, it was found that there is potential to save up to $200 per patient if a proper screening is performed when a patient enters the system for outpatient surgery. Based on the number of patients that APC currently sees, this represents a savings of almost $70,000 per month.

Potential cost savings (or avoidance of lost revenues) and improved outcomes due to anesthesia pre-operative clinic assessment are not unique to our particular case study. Correll et al. (2006) documented the value of preoperative clinic visits on reducing operating room delays and cancellations at Brigham and Women's Hospital in Boston, Massachusetts. This study valued lost operating room time due to cancellation or delays at several hundreds of dollars per hour. van Klei et al. (2002) conducted a large sample study on elective adult inpatients pre and post introduction of an anesthesia preoperative assessment clinic at University Medical Center of Utrecht in the Netherlands. They showed a significant relative reduction in surgery cancellations for medical reasons of 53% and cited a number of other studies that showed relative reduction in surgical cancellations due to advanced anesthesia preoperative assessment ranging from 20% to 88%. In addition, van Klei et al. (2002) also reported a significant reduction in total hospital length of stay. Ferschl et al.

(2005) conducted a retrospective study on over 6000 surgical cases from a six month time period at the University of Chicago Hospitals. They found a significant reduction in cancellation rates (at least 50%) and delay rates for patients receiving advanced anesthesia preoperative assessment. Edward et al. (2008) conducted an analysis on patients undergoing elective non-cardiac surgery over a six year time period at Brigham and Women's Hospital in Boston, Massachusetts. They found a significant reduction in consultations when proper evidence-based protocols were used in anesthesia preoperative clinic assessment.

2.3.1 The Original System

Figure 2.1 outlines the original system. The 40-60% referral/non-referral split to APC is determined by the surgeons at the clinics of origin. This triage is supposed to be guided by the underlying medical condition of the surgical patient; those with more comorbid conditions would be referred to the APC. The ASA Physical Status Classification System (ASA (2013)) correlates with the degree and severity of coexisting conditions. However, this classification is not formally assigned to a patient until he or she is evaluated by an anesthesiologist, either in the APC or on the day of surgery. Most patients in ASA classes 1 and 2 (i.e., patients who are in general good health with minor medical problems) do not need an in-clinic APC consultation. Those patients classified as ASA 3 and 4 (i.e., patients with more complex medical problems) should be referred for an in-clinic consultation at APC.

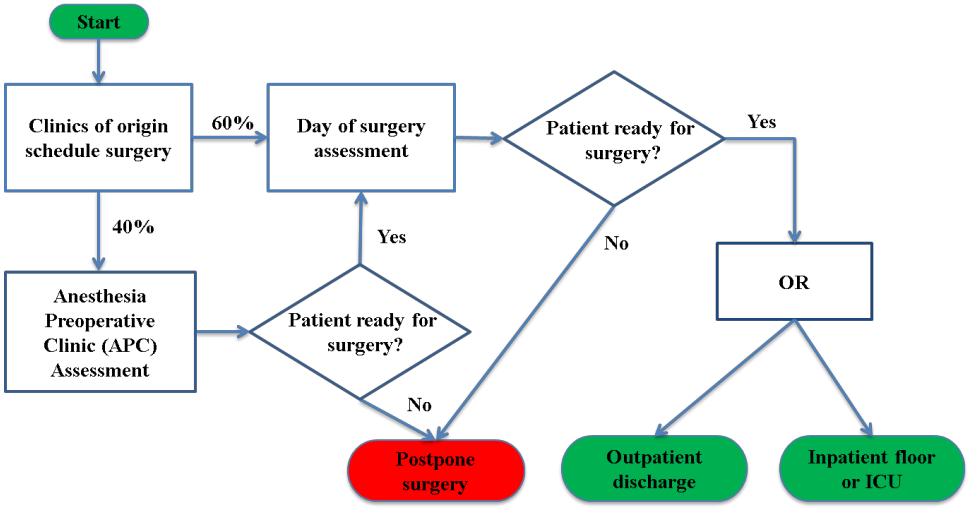

Figure 2.1: An overview of the original outpatient surgery system

Having two assessments of each patient's state of health (one at the surgeon's clinic and one at APC or day of surgery assessment) allowed us to estimate the proportion of ASA 3 and 4 patients not sent to APC (an incorrect decision that, if corrected, would increase APC's in-clinic demand) and the proportion of ASA 1 and 2 patients sent to APC (an incorrect decision that, if corrected, would reduce APC's in-clinic demand). It also allowed us to estimate of the difference between these two quantities to determine what proportion of patients that APC should expect to see in-clinic if the triage were done properly according to the ASA classification.

Figure 2.2 shows the results of how well the surgical clinics are performing this triage in the original system. These data were collected on 370 patients over 11 days from February through May 2013 in the outpatient surgery holding area on the day of surgery. The data collection days were determined by the availability of person-

nel to collect the data and to ensure that all days of the week were observed. On data collection days, all outpatients were included in the sample. These data reveal inadequate triage by the surgical clinics is a problem. Of the 148 APC patients, 32 (almost 22%) were classified as ASA 1 and 2. About 36% (81 out of 222) of the non-APC patients were classified as ASA 3 and 4. If all patients were triaged to APC based on their ASA 3 and 4 status, then, from this sample, we estimate that the APC in-clinic consultation demand would grow by about 33% (a net increase in demand of 49).

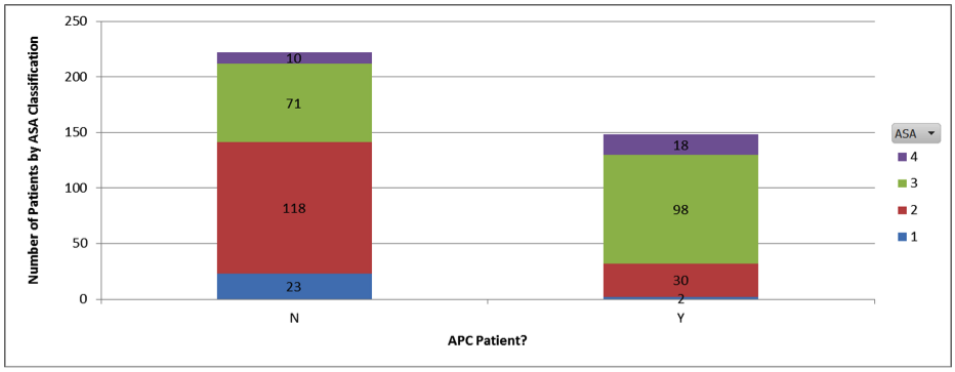

Figure 2.2: ASA Classification of APC and Non-APC Patients

More rigorously, this problem can be formulated as a statistical confidence interval for the difference between two correlated proportions (Lloyd (1990)). It is also referred to as "the contrast between two proportions based on individually paired data" (Newcombe (1998)). Using Newcombe's approach (see Appendix A.1), our data yields an approximate 95% confidence interval of (0.078, 0.187) for the difference between the proportion of ASA 3 and 4 patients not sent to APC and the proportion of ASA 1 and 2 patients sent to APC. Hence, with 95% confidence, the

proportion of APC patients would increase between 0.078 and 0.187 over the proportion of 0.4 in the original system (represents an increase of between 19.5% and 46.75%) if the patients were properly triaged.

As a side benefit of the misclassification of patients shown in Figure 2.2, we were able to assess the benefit of APC on potentially preventable day of surgery delays. While this has been firmly established in the literature by Ferschl et al. (2005) in a more extensive study, we thought that it would be worthwhile to test this hypothesis in our own context. Of the 370 patients, more extensive information was collected on 149 which included whether or not these patients experienced potentially preventable day of surgery delays. Table 2.1 shows results on all 79 ASA 3 and 4 patients in this data set. These are the patients that were either seen at APC or should have been seen at APC (referred to as "non-APC"). Based on the theory of a randomized statistical test, the likelihood of the results in Table 2.1 or something more extreme (i.e., APC having no potentially preventable delays and Non-APC accounting for 7 potential delays) occurring randomly is 0.0434 (see Appendix A.2 for likelihood calculations). Hence, there is strong evidence to suggest that APC is effective at reducing potentially preventable delays on the day of surgery. Furthermore, as Ferschl et al. (2005) point out, this result is conservative because patient assignment was not done randomly, rather "sicker patients" were assigned to APC.

	Delayed (Potentially Preventable)	Not Delayed	Total
APC	1	40	41
Non-APC	6	32	38

Table 2.1: Potentially preventable day of surgery delays for APC and Non-APC ASA 3 and 4 Patients

Since APC plays a key role in PCSH, we examine its workflows in more detail. Figure 2.3 contains a patient flow process mapping of the original APC. When a patient arrives at the clinic, she checks in and registers with a clerk. A registered nurse (RN) checks patient vitals and collects other necessary information. The provider examination is conducted by a resident physician or nurse practitioner with faculty oversight. Finally, the patient is discharged. Regarding resources, APC has one clerk, an RN, two residents, a nurse practitioner, and one attending anesthesiologist faculty member.

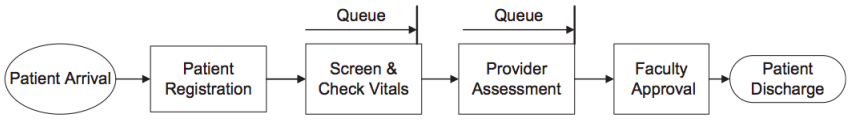

Figure 2.3: APC patient flow process mapping

We conducted an observational study of the APC process during a five-week time period (June 28, 2012 to August 1, 2012) that included 356 patients. When combined with historical data from clinic scheduling and patient check-in records, clinic referral information, and monthly patient visits, an analysis of these data revealed several things. First, providers were the limiting resource (process bottle-

neck). More specifically, each provider spent, on average, 39 minutes with each patient and roughly another 15-18 minutes (on average) attending to informational issues immediately after each patient's departure (total: 54-57 minutes). Since there were 3 providers, the process capacity was one patient every 18-19 minutes. Patients arrived at roughly one patient every 20 to 23 minutes. Therefore, provider utilization was in the 80 to 95% range. As a result, the APC process was running close to capacity, and experiencing high congestion.

Second, there was high variability in patient arrivals. We found substantial patient arrival rate variation both within and across days. Additionally, historical data showed a 15.45% average no-show rate and significant deviation between actual and scheduled arrival times.

We also found high variability in the amount of information that needed to be collected by providers for patients during an APC visit. Not all patients had medical records sent in advance, and not all patients knew sufficient details of their medical history to inform an anesthesia assessment. This had significant impact on the provider assessment time and the duration of the clinic visit. Because wait times can be a source of patient dissatisfaction (Hepner et al. (2004)), the long clinic visit times we observed were of concern to APC.

Each patient observed in the five-week study was rated by the faculty anesthesiologist on amount of information deficiency (low, medium, or high). A low rated patient required only straightforward updating of his/her electronic medical record (EMR). If, in addition, a patient required information from an external source (e.g., primary care provider or specialist) that required one phone call and some EMR

searching, then this patient was considered medium. A high rated patient needed information from multiple external sources requiring more than one phone call and extensive searching of EMR databases. Figure 2.4 provides a summary of these results along with the impact of information deficiency on average provider assessment time for patients. Fifty-nine percent of APC patients have medium or high information deficiency. Average provider assessment time nearly doubled from low to high information deficiency. This is an example of the so-called "information hangover" discussed by Lahiri and Seidmann (2012).

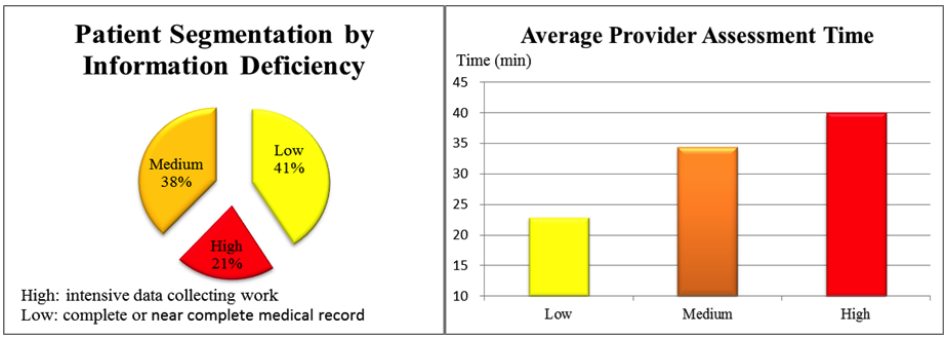

Figure 2.4: Patient segmentation by information deficiency and its impact on provider assessment time

In the five-week study, the faculty anesthesiologist also rated each patient on surgical and patient complexity. Surgical complexity was assessed as low, medium, or high based on the complexity of the surgical procedure (as described in Eagle et al. (2002)). Additionally, patient complexity was assessed as low (patient in relatively good health), medium, or high (patient in relatively poor health). This assessment was guided by the ASA Physical Status Classification System (ASA (2013)). In Sections 2.4 and 2.5, we will show that these factors also impact provider times.

Combined, all these issues posed a major challenge for APC and the implementation of the PCSH model. We addressed this challenge by redefining APC's role as system coordinator and then simulating to determine how it could fulfill this role.

2.3.2 The New System

Figure 2.5 shows a diagram of the new system in which APC acts as the system-wide coordinator of the PCSH by assessing all patients needing surgery using either a clinic visit for the more complicated patients (ASA 3 and 4) or a telephone screening for the healthier patients (ASA 1 and 2). Figure 2.6 illustrates key features of the APC-coordinated PCSH model. It relies on a screening tool that is completed by each patient and an RN who navigates each patient's case through the entire process. Patients sign a release of information consent for the APC to obtain the necessary medical records from providers outside the UTM / UHS system. Using the results from the screening tool, the RN navigator determines whether a patient needs to make an APC visit appointment or can be screened over the telephone prior to surgery. The decision is based upon the patient's medical history. The RN navigator confirms that appropriate tests are ordered and initiates retrieval of information for each patient. These steps are done in advance by the RN to increase the percentage of low information deficiency patients seen in-clinic at APC. Telephone screening is conducted by the RN navigator, saving clinic provider time for those patients that really need to be seen in APC. The RN navigator completes the documentation in each patient's record, ensuring that the patient is ready for surgery.

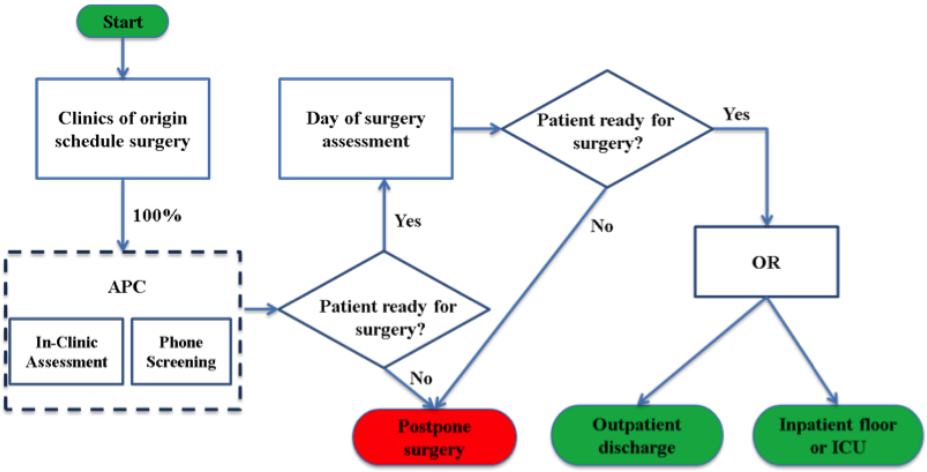

Figure 2.5: An overview of the new outpatient surgery system

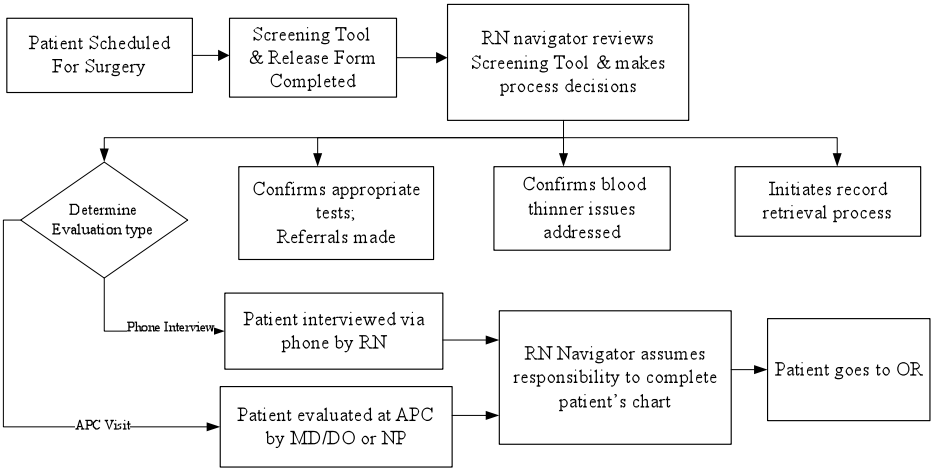

Figure 2.6: Key features of an APC-coordinated PCSH model

Transitioning the APC into its role as system coordinator required the development of a screening tool. Additionally, it required redeploying the RN from checking vitals to the navigator role depicted in Figure 2.6. APC replaced the RN at the

check vitals step with two Medical Technicians (MTs), which was more economical than hiring another RN. Two MTs were hired to have extra personnel available to perform additional tasks such as conducting electrocardiograms and to ensure that the check vitals step did not become a bottleneck.

2.4. The APC Simulation

Using the observational and historical data along with the process mapping information described in Section 2.3.1, we constructed a process simulation of the original APC system using the Arena simulation software (Kelton et al. (2010)). A preliminary version of this model appears in Morrice et al. (2013). We chose not to use queuing theory in our analysis like Zonderland et al. (2009) because patients arriving on time and steady state analysis did not hold (even approximately) in APC. Additionally, simulation allowed us the flexibility to explicitly model patient complexity, surgical complexity, and information deficiency for provider times.

2.4.1 The Simulation Model

Figure 2.7 contains a screen capture of the simulation model of the APC process depicted in Figure 2.3. Patient arrival times were assigned as deviations off scheduled appointment times. Different types of patients were generated based on information deficiency, patient complexity, and surgical complexity. Wrap-up is the time providers spend gathering information entering data into the EMR after the patient is discharged. The provider assessment and wrap-up process steps in the simulation account for faculty approval. The simulation also models the staffing

and scheduling requirements for the resources (details not shown in the diagram).

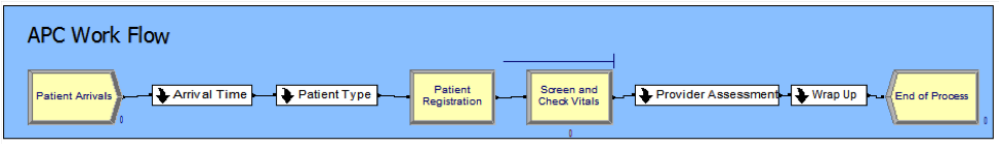

Figure 2.7: Arena simulation model of APC

Input models for patient arrivals and processing times in the simulation were estimated from the observational and historical data using the Arena Input Analyzer (Arena (2013)), EasyFit (EasyFit (2013)), StatTools (Palisade (2013)) and STATA (STATA (2013)). Table 2.2 provides a summary of these models. Provider assessment and wrap-up times are represented as functions of the information deficiency in the simulation model, which is to be expected based on Figure 2.4. Provider assessment time was also found to be a function of the surgical and patient complexity.

# of Patients Per Day	Arrival Process	Registration	Nurse Assessment	Provider Assessment	Provider Wrap-up
Discrete Empirical	Johnson SU (deviation off scheduled times)	Lognormal	Erlang + Triangular	Regression (function of surgical and patient complexity, and information deficiency)	Regression (function of information deficiency)

Table 2.2: Models for patient arrivals and processing times in the APC simulation

We found that patient registration at the clinic followed a Lognormal distribution. We used a Discrete Empirical distribution for the number of patients per day because no parametric distribution was found to provide a good fit to the data. An Erlang distribution was used for nurse assessment. However, after the fact, we found "gaps" in the data in which the nurse was summoned to perform other duties.

These gaps were mostly explained by requests for the nurse to conduct electrocardiogram (EKG) tests on the patients. Since there was only one nurse we were able to glean some of these times from the data and model them using a Triangular distribution. We will describe the models used for the other three processes in more detail.

2.4.1.1 Model for Patient Arrivals

To model arrivals, we used the deviation time between arrival and scheduled times. The Johnson SU performed the best among 36 candidate distributions in terms of Kolmogorov Smirnov, Anderson Darling and Chi-squared tests. We believe Johnson SU is appropriate for patient arrivals for two reasons: (1) it allows for heavier tails which correspond to extreme deviation values due to logistical challenges faced by the patient population (e.g., inner city transportation issues); and (2) it accommodates for a high peak at zero since many patients arrive punctually. Figure 2.8 illustrates the fit of the Johnson SU distribution using histogram/density function and P-P plots. Incidentally, our findings are in line with Alexopoulos et al. (2008); Rohleder et al. (2011). Similar to these studies, we report on the superiority of the Johnson SU distribution over the Normal distribution for APC arrivals in Morrice et al. (2013).

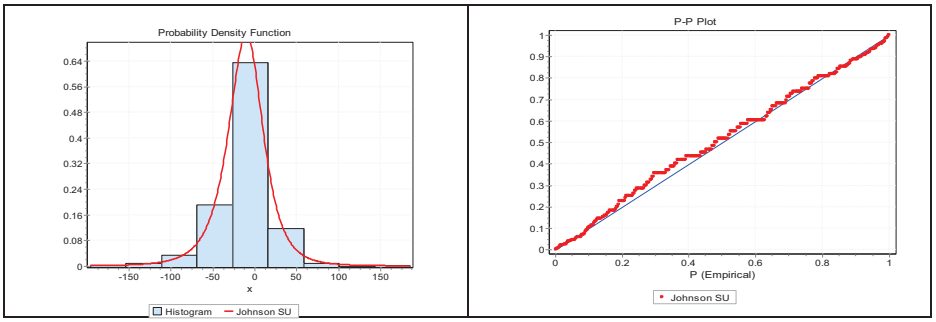

Figure 2.8: Arrival deviation fit in Johnson SU distribution

2.4.1.2 Model for Provider Assessment

For provider assessment time, we regressed the natural log of provider assessment time on dummy variables associated with surgical complexity (S), patient complexity (P) and information deficiency (I) using backward regression analysis. The regression model with an adjusted $R^2 = 0.4044$ is given in Equation (2.1). The "low" dummies (with subscript "L") were left out of the regression. The subscripts "M" and "H" refer to the medium and high dummy variables, respectively. All variables, except for S_H (p-value = 0.245), were statistically significant at the 5% level. The variable I_H has the largest coefficient in the regression model. It shows that high information deficiency plays a greater role in longer provider assessment times than high surgical complexity and poor physical health. We elected to leave S_H in the regression model because it made practical and theoretical sense to do so. Also including this variable provided a slight improvement in the Adjusted R^2 level. It is important to note that we did try including first-order interaction terms, but the few that were found to be statistically significant in the model were collinear

with the main effects and therefore not worth retaining.

$$\begin{aligned} Ln(\textit{Provider Assessment Time}) &= 3.06 + 0.37P_H + 0.25P_M + 0.1S_H + 0.11S_M \\ &\quad +0.56I_H + 0.21I_M + \epsilon \end{aligned} \tag{2.1}$$

We ran several diagnostic tests on the residuals. Neither the chi-square test nor the Lilliefors test rejected normality. No significant deviations from the assumption of statistical independence were detected by the autocorrelations and runs tests. Finally, a residuals versus fitted plot showed no evidence of non-linearity or heteroscedasticity. It is important to note that the residuals of the regression are log-normal. Lognormal service times have been found to be quite common in other studies (Cayirli et al. (2006); Rohleder et al. (2011)).

2.4.1.3 Model for Provider Wrap-up Times

In the summer 2012 five-week observational study of APC, we did not initially collect data on the time taken by providers for additional information gathering and EMR data entry after the patient is discharged (provider wrap-up). Consequently, we did not capture all the time a provider spent on each case and our simulation queuing statistics did not match the observed data (the simulation queue lengths were much shorter than what was being observed in the real system). Once we realized this, we conducted a focused observational study on provider wrap-up times and collected 107 observations on eight days in December 2012 through February 2013. The eight days were selected based on the availability of an observer to col-

lect data, the convenience of the clinic, and to ensure that all days of the week were observed. Wrap-up times were regressed on S, P and I. From backward regression, only I_H and I_M remained in the model (see Equation (2.2)). The Adjusted R^2 for this model is 0.3053. Again, the coefficients highlight the impact of information deficiency.

$$Wrap - up\ Time = 11.46 + 10.67I_H + 5.13I_M + \epsilon \tag{2.2}$$

Based on the chi-square and Lilliefors test, the residuals in the regression model in Equation (2.2) did not satisfy the normality assumption. Instead, a Beta distribution was found to fit the residuals using the Arena Input Analyzer.

2.4.1.4 Additional Comments on Provider Assessment and Wrap-up Times

We did test the effect of different providers on both assessment and wrap-up times and it was found to be not significant on either of these times. This is likely to due to the fact that providers' (residents and nurse practitioner) work has to be reviewed by the attending faculty member who governs the pace of the providers. We could not fully assess the impact of variance in faculty wrap-up time because one faculty member provides the wrap-up care for the majority of the patients.

2.4.2 Model Validation

Table 2.3 contains a comparison of the observed processing times from the summer 2012 five-week observational study and the statistics generated by the simulation. The simulation was run for 600 days (over two years of clinic time). Since

the number of patient arrivals varied from one day to the next, 600 was chosen to ensure enough days were simulated even for the lowest probability number of patients per day. The results indicated that for most of the processes, the simulation produced results that were not statistically distinguishable from the observed data. The one statistic of concern was the time for extra nurse duties. These data had to be crudely approximated from gaps in the data so this was not entirely surprising. We did not launch a separate study to get more data on the RN because the nurse was not the bottleneck in the system, and she was replaced by the MTs in the new system. Details of the MT input distributions are provided in the next section.

	Observation	**Simulation**	
Process	**Mean**	**Mean**	**96% CI Half-width**
Number of Pateints per Day	15.85	15.79	0.22
Registration	8.71	8.60	0.14
Nurse Queuing Time	20.67	21.39	0.64
Nurse Assessment Time	12.42	12.35	0.09
Time for Exra Nurse Duties	5.57	2.7	0.01
Provider Queuing Time	15.85	16.61	0.94
Provider Assessment Time	38.42	38.22	0.37
Provider Wrap-up Time	15.16	15.71	0.17
Patient Waiting Time in System	96.98	97.17	1.31

Table 2.3: Simulation validation statistics

2.5. Simulation Analysis for the APC-coordinated PCSH Model

Using the simulation, we designed an experiment of the new system described in Section 2.3.2. Since the MTs replaced the nurse at the check vitals step in the APC process in the new system, we conducted an observational study of the MTs perfor-

mance in APC from April 19 to June 19, 2013. This yielded 467 medical technician assessment times to which we fit a gamma distribution with a mean of approximately 18 minutes using Arena Input Analyzer. Additionally, we were able to estimate that 9% of patients required an EKG and the time taken for an EKG was found to follow a triangular distribution with mean close to eight minutes. We updated the simulation model in Figure 2.7 to reflect these changes and the resultant model was used in the following simulation experiments.

To simplify the experimental design, we simulated only heavy patient load days on which there were zero no-shows. This was satisfactory for our analysis because UHS administrators require APC to have the capabilities to handle such days. To determine a heavy patient load, we used results from the observational study depicted in Figure 2.2. From this, we estimated that, on average, about 53% of the UHS outpatient surgery patients were ASA 3 or 4 and would need to be seen in the clinic at APC. From historical records of the daily number of cases at UHS outpatient surgery from 07/25/2011 to 03/15/2013, we determined that 25 patients/day represented the 99th percentile of the number of patients needed to be seen by APC in the clinic (i.e., ASA 3 and 4 patients) on any given day.

As a second simplification of the experimental design, we split the analysis into two phases. In the first phase, we considered the main factors to be information deficiency and patient complexity. Note: we did not consider surgical complexity, because it had less impact than the other two factors (see Equations (2.1) and (2.2)) on provider assessment and wrap-up times. Additionally, surgical complexity was not expected to change. As a result, there was no plan to use this factor in the re-

design of the system. The design points of interest from the first phase were carried to the second phase where we considered them in conjunction with three other factors that could potentially improve performance of the system: patient scheduling, the patient arrival window around scheduled appointment times, and MT starting times.

2.5.1 Phase One Experiments

For patient complexity, we considered two levels in the experiment design: level one (denoted by $P_L\_P_M\_P_H$) is the baseline case of 21% P_H, 60% P_M, and 19% P_L patients from the summer 2012 five-week observational study; and level two ($P_M\_P_H$) where P_L drops out, and APC sees only medium and high complexity cases in the same relative proportions (i.e., 25.9% P_H, 74.1% P_M). The latter case represents the APC-coordinated PCSH model where the screening tool and the RN navigator are perfectly effective and APC sees all the most complicated cases in clinic, as it should. We did not consider any other cases because there is no reason to believe that the APC-coordinated PCSH model would result in more low complexity patients. Furthermore, the objective of this study was to assess if APC could handle the increased complexity patient load.

Regarding information deficiency, we considered three levels: level one ($I_L\_I_M\_I_H$) which is the baseline case from the summer 2012 five-week observational study (41% I_L, 38% I_M, 21% I_H – see Figure 2.3); level two ($I_L\_I_M$) where the APC-coordinated PCSH model is effective at eliminating all high information deficiencies, but the low and medium information deficiencies remain in the same relative proportions (i.e.,

51.9% I_L, 48.1% I_M); and level three (I_L) where the APC-coordinated PCSH model is effective at eliminating all high and medium information deficiencies and only low information deficiency patients show up in the APC clinic. Again, we did not consider any other cases because we had no reason to believe that the APC-coordinated PCSH model would result in higher information deficiencies.

Table 2.4 provides a summary of the simulation results for the six scenarios. The results were based on 600 simulated days with 25 patients per day. For all scenarios, clinic staff started at 7:00 am. In addition, the first patients were scheduled at 7:30 am, although they could enter any time after 7:00 am. This was designed to mimic reality where patients often show up early at the beginning of the day. The same appointment schedule was used for all six scenarios. More details on appointment schedules will be discussed in the second phase of the experimental design. Note: i) 95% refers to the 95th percentile of the distribution, ii) HW refers to a 95% confidence interval half-width estimated by Arena, and iii) Clinic Session Length is the length of time the clinic is open each day after 7:00 am. The other statistics are self-explanatory.

Scenario	(Information Deficiency Level, Patient complexity Level)											
	(I_L-I_M-I_H, P_L-P_M-P_H)		(I_L-I_M-I_H, P_M-P_H)		(I_L-I_M, P_L-P_M-P_H)		(I_L-I_M, P_M-P_H)		(I_L, P_L-P_M-P_H)		(I_L, P_M-P_H)	
Statistics (times are in minutes)	Mean (95%)	HW	Mean (95%)	HW	Mean (95%)	HW	Mean (95%)	HW	Mean (95%)	HW	Mean (95%)	HW
Patient Total Time in Clinic	86.17 (144.07)	<1.20	89.29 (148.06)	<1.17	71.72 (113.47)	<0.63	75.06 (119.19)	<0.72	64.19 (98.73)	<0.46	65.91 (101.86)	<0.47
Clinic Session Length	554.93 (608.96)	2.71	559.84 (618.79)	<2.66	538.00 (580.66)	<1.97	539.78 (584.21)	<2.33	528.51 (576.19)	<2.20	528.42 (572.58)	2.08
Provider Utilization	80.87%	NA	82.31%	NA	74.38%	NA	76.27%	NA	66.21%	NA	68.12%	NA
MT Process Time	17.94	<0.09	17.93	<0.09	17.91	<0.09	17.95	<0.09	18.03	<0.10	17.93	<0.10
Provider Assessment Time	38.07	<0.29	39.65	<0.27	33.78	<0.22	35.44	<0.23	30.39	<0.19	31.82	<0.19
Provider Wrap Up Time	15.73	<0.13	15.58	<0.12	13.97	<0.12	13.87	<0.11	11.51	<0.11	11.28	<0.11
Queuing Time for MT	1.77	<0.09	1.64	<0.09	1.82	<0.09	1.76	<0.09	1.89	<0.10	1.68	<0.09
Queuing Time for Provider	19.68	<0.97	21.49	<0.95	9.48	<0.45	11.30	<0.53	5.23	<0.28	5.82	<0.31

Table 2.4: Comparison of six scenarios for information deficiency and patient complexity

Patient total time in clinic was the primary statistic of interest to the clinic personnel and hospital administration. With the changes, their target for this measure was a 60 minute average. The results in Table 2.4 indicate that lowering information deficiency, whether moderately or substantially, yields significantly lower total patient times which should improve patient satisfaction, an important component of a patient-centered model. The reduction is attributable to lower provider assessment and wrap-up times along with the attendant reduction in provider queuing time. While none of the scenarios achieved the desired target of 60 minutes, the scenarios with low patient information came within five to six minutes, even with more complex patients.

We conducted a more thorough analysis of the impact of information deficiency and patient complexity on total patient time using regression. To satisfy the standard regression assumptions of normality, homoscadasticity, linearity and indepen-

dence, we:

1. Used non-overlapping random number streams in Arena to ensure that the runs at each design point and across design points were statistically independent.

2. Formed batches of 30 replications, yielding 20 batch means at each design point to approximate normality (Law and Kelton (2010), Chapter 9).

3. Conducted weighted least squares since the residuals across design points were found to exhibit heteroscadasticity. The weights were estimated from the standard error of the batch means at each design point (Judge et al. (1988)).

The regression results for total patient time regressed on levels of information deficiency and patient complexity shown in Table 2.5 were generated by StatTools (Palisade (2013)). We confirmed normality (chi-square and Lilliefor's tests), homoscedasticity and linearity (visual inspection of the residual versus fitted plot), and statistical independence (autocorrelation test and runs test for randomness). The regression results confirm the high impact of lower information deficiency on the total patient time. They also demonstrate that better triage resulting in more complex patients in clinic will increase total patient time. What is most surprising about the results in Table 2.5 is the significant interaction term. It indicates that if APC can achieve low information deficiency and the proper triage of patients, a reinforcing improvement on total patient time occurs. One possible explanation for this reinforcing effect is improved consistency of provider times resulting from better information and a better triage. While in absolute terms the interaction effect

seems small, it offsets almost half of the increase resulting from seeing more complex patients in the clinic. Incidentally, the interaction term of $I_{L\_}I_M \times P_{M\_}P_H$ was not statistically significant (and hence, not included in the model). This further underscores the importance of getting the best patient information possible prior to a clinic visit.

	Model	
	Adjusted-$R^2 = 0.9940$	
Regression Model Term	Coefficient	P-value
Constant	86.25	< 0.0001
Dummy ($I_{L\_}I_M$)	-14.49	< 0.0001
Dummy (I_L)	-22.14	< 0.0001
Dummy ($P_{M\_}P_H$)	3.25	< 0.0001
Interaction ($I_L \times P_{M\_}P_H$)	-1.36	0.0274

Table 2.5: Total patient time regressed on levels of information deficiency and patient complexity

Two other measures in Table 2.4 were of secondary interest to clinic management: clinic session length and provider utilization. Since both showed improvement with lower information deficiency that more than offset any degradation associated with seeing more complex patients in the clinic (see Figure 2.4), clinic management were satisfied with these results.

2.5.2 Phase Two Experiments

In this second phase of experiments, we looked for ways to improve the desired designs for the APC-coordinated PCSH model: the design that is effective at eliminating all high information deficiencies and does a proper triage ($I_{L\_}I_M$, $P_{M\_}P_H$),

and the design that is effective at eliminating all high and medium information deficiencies and does a proper triage (I_L, $P_M\_P_H$). We bring the first design forward, rather than the second only, to gain some understanding of system performance if APC is not completely effective at eliminating all high and medium information deficiencies. As in the first phase, results for all scenarios are based on 600 simulated days with 25 patients per day, and clinic staff report for duty at 7:00 am, unless stated otherwise.

Table 2.6 contains the current patient schedule used by APC along with the patients' schedules generated by several policies considered by Cayirli et al. (2006) and Millhiser et al. (2012) including an individual-block/fixed-interval (IBFI) rule, a variant of Bailey's Rule (Bailey (1952)) with 4 patients starting at the beginning of the day (one more than the number of providers) (4BEG), a two-block/fixed interval (2BFI) rule, and an individual-block/variable interval rule that results in "dome-shaped" appointment intervals (DOME). Table 7 contains the simulation results for different scheduling policies on scenario ($I_L\_I_M$, $P_M\_P_H$). Notice that IBFI, 4BEG, and 2BFI perform worse on patient total time in clinic but better on clinic session length. While the DOME policy is on the borderline of being statistically better than the current policy on patient total time, the former is definitely statistically inferior to the latter on clinic session length. Given that 2BFI was better on patient total time and at least as good on clinic session length as IBFI and 4BEG, and DOME showed some potential to be better than the current policy on patient total time for Scenario ($I_L\_I_M$, $P_M\_P_H$), we applied 2BFI and DOME to (I_L, $P_M\_P_H$) scenario. Table 2.8 contains these results. On the primary performance measure, patient total time in clinic,

it appears nothing would be gained by switching from the current patient scheduling policy to one of these other policies. However, if clinic management were willing to make a trade-off between patient total time and clinic session length, then 2BFI should be given serious consideration.

Policy												
Current	Time	7:30	8:30	9:00	9:30	10:00	10:30	11:30	13:00	13:30	14:00	14:30
	# of patients	4	2	2	2	2	2	2	2	2	2	1
IBFI	Time	7:30	Every 15 minutes till 11:30						13:00	Every 15 minutes till 14:00		
	# of patients	3	1 patient per session						2	1 patient per session		
4BEG	Time	7:30	Every 15 minutes till 11:00						13:00	Every 15 minutes till 14:00		
	# of patients	4	1 patient per session						3	1 patient per session		
2BFI	Time	7:30	8:00	8:30	9:00	9:30	10:00	10:30	11:30	13:00	13:30	14:00
	# of patients	3	2	2	2	2	2	2	2	3	2	1
DOME	Time	7:30	7:40	7:50	8:05	8:45	9:10	9:40	10:05	1-:30	10:50	11:05
	# of patients	3	1	1	1	1	1	1	1	1	1	1
	Time	11:20	11:25	13:00	13:10	13:20	13:35	13:55	14:15	14:40		
	# of patients	1	1	3	1	1	1	1	1	1		

Table 2.6: Patient schedules for current APC policy and several policies from the literature

	Policies									
	Current		IBFI		4BEG		2BFI		DOME	
Statistics	Mean (95%)	Hw	Mean (95%)	Hw	Mean (95%)	Hw	Mean (95%)	Hw	Mean (95%)	Hw
Patient Total Time in Clinic	75.06 (119.19)	<0.72	83.58 (136.09)	<1.02	85.55 (138.95)	<0.96	81.05 (129.31)	<0.95	73.60 (115.48)	<0.57
Clinic Session Length	539.78 (584.21)	<2.33	515.08 (555.59)	<2.09	521.07 (566.87)	<2.17	512.51 (554.63)	<2.24	553.82 (609.40)	<2.37
Provider Utilization	76.27%	NA	80.28%	NA	79.14%	NA	80.10%	NA	74.58%	NA

Table 2.7: Simulation results from different scheduling policies on scenario ($I_L\_I_M$, $P_M\_P_H$)

	Patient Scheduling Policies					
	Current		2BFI		DOME	
Statistics	Mean (95%)	Hw	Mean (95%)	Hw	Mean (95%)	Hw
Patient Total Time in Clinic	65.91 (101.86)	<0.47	69.20 (106.90)	<0.59	66.04 (101.00)	<0.40
Clinic Session Length	528.42 (572.58)	2.08	499.73 (549.57)	<2.34	540.92 (597.93)	<2.42
Provider Utilization	68.12%	NA	72.37%	NA	67.04%	NA

Table 2.8: Simulation results for different scheduling policies on scenario (I_L, P_M_P_H)

Table 2.9 contains results from assuming a 30-minute patient arrival window around scheduled appointment times. This was constructed in the simulation by simply truncating the Johnson distribution representing deviation off a patient's scheduled time depicted in Figure 2.8. While the results suggest that such an arrival window might be beneficial, they are not statistically significant. Hence, these results provide no compelling reason to pursue such a policy which can be challenging to enforce in practice. This is particularly important for the APC, since many of its patients are from an under-served inner city population in San Antonio who must rely on public transportation schedules that may not synchronize well with the clinic schedule.

	Policies for Scenario (I_L-I_M, P_M-P_H)				Policies for Scenario (I_L, P_M-P_H)			
	Current		30-min Cut-off		Current		30-min Cut-off	
Statistics	Mean (95%)	Hw	Mean (95%)	Hw	Mean (95%)	Hw	Mean (95%)	Hw
Patient Total Time in Clinic	75.06 (119.19)	<0.72	74.33 (117.13)	<0.67	65.91 (101.86)	<0.47	65.05 (97.18)	<0.41
Clinic Session Length	539.78 (584.21)	<2.33	537.07 (572.38)	<1.60	528.42 (572.58)	2.08	525.21 (556.75)	<1.44
Provider Utilization	76.27%	NA	76.65%	NA	68.12%	NA	68.61%	NA

Table 2.9: Simulation results assuming a 30-minute patient arrival window around scheduled appointment times

The last scenario we considered was offsetting the starting times of the two MTs by 30 minutes, having one start at 7:00 am and the other start at 7:30 am. The potential advantage of this policy is having an MT available later in the day to service late arriving patients (Table 2.10). The MT overtime measures the percentage of simulation runs in which we found an MT would be retained past an eight hour shift to process a patient. By having one MT start 30 minutes later than the other, we estimated that the probability of having MT overtime decreased by almost 5% in both the (I_L_I_M, P_M_P_H) and (I_L, P_M_P_H) scenarios. Since this strategy did not yield statistically worse results than the current policy of having both MTs start at 7:00 am on the other performance measures, our simulation results indicate that this is a strategy worth pursuing.

	Policies for Scenario (I_L-I_M, P_M-P_H)				Policies for Scenario (I_L, P_M-P_H)			
Statistics	Current Percentage		MT Offset Percentage		Current Percentage		MT Offset Percentage	
MT Overtime	5.33%		0.00%		4.50%		0.83%	
Statistics	Mean (95%)	Hw	Mean (95%)	Hw	Mean (95%)	Hw	Mean (95%)	Hw
Patient Total Time in Clinic	75.06 (119.19)	<0.72	74.33 (117.13)	<0.67	65.91 (101.86)	<0.47	65.05 (97.18)	<0.41
Clinic Session Length	539.78 (584.21)	<2.33	537.07 (572.38)	<1.60	528.42 (572.58)	<2.08	525.21 (556.75)	<1.44
Provider Utilization	76.27%	NA	76.65%	NA	68.12%	NA	68.61%	NA

Table 2.10: Simulation results when one MT starts at 7:00 am and the other starts at 7:30 am

2.6. Discussion and Insights

We have shown a way to implement the ASA vision of a PSH that will enable significant improvements in patient care, OR room utilization, and diagnostic testing. It is important to note that the addition of two MTs increased the APC clinic

budget by about $50,000 per year. Given the potential benefits for OR utilization and patient testing, this represents about a 20-fold return on investment.

When combined with patient complexity and surgical complexity, information deficiency poses great challenges to the realization of the PCSH, both in terms of optimal clinical management of the patient and logistical operation of the APC. However, with better coordination, staffing, and the use of a screening tool to assess patient history and potential information needs prior to the time of the visit, we have shown that APC is well-positioned to serve as the coordinator and information integrator of the PCSH.

In addition to mitigating information deficiency, the APC-coordinated PCSH must be designed to properly triage patients. We approached this issue by utilizing an RN who is able to follow a clinical protocol to assign patients to either a clinic or phone visit. The redeployment of the RN and introduction of the MTs enables better utilization of all staff, allowing the RN to make clinical assessments, allowing the MTs to obtain vitals and more basic information, and allowing the physicians and nurse practitioner to spend a greater proportion of time seeing patients. This is both more efficient and cost-effective. Moreover, using simulation we have shown that proper triage coupled with low information deficiency results in reinforcing improvements on patient times in the clinic.

Additional analysis comparing the current scheduling policy with rules proposed in the literature revealed that on the primary performance measure of patient total time in clinic, nothing would be gained by switching from the current policy. However, if clinic management were willing to make a trade-off between patient to-

tal time and clinic session length, then 2BFI should be given serious consideration.

Using a simulation approach to develop the PCSH model in the APC also allowed for more efficient implementation. Rather than making process decisions and staffing changes through trial and error, we were able to make more informed decisions regarding what processes of care to pursue. This prevented inefficiencies both in terms of avoiding pursuing ineffective models and of implementing the actual processes – inefficiencies that could have real cost in terms of clinic staff and patient care. When these results were presented to top UHS hospital management, APC was given approval to implement the model in Figures 2.5 and 2.6. The necessary adjustments were made to the clinic staff to implement this process and pilot the screening tool.

According to the ASA, the PSH model will provide an opportunity to utilize a bundled pay model rather than the traditional fee-for-service. The thought is that quality and efficiency will be improved and cost will go down. Cost savings will result from fewer mistakes and accidents, better care, standardized testing (eliminating duplication of services), reduced cancellations and delays on day of surgery, less variation in overall care (including standardization of materials, implants, medications, etc.), and reduced hospital readmissions. In summary, the model should improve patient outcome per dollar expended.

2.7. Summary and Conclusions

In this study, we have provided a PCSH model for outpatient surgery. The key to establishing a PCSH was the insight that APC could serve as system coordinator. Without such coordination, we have shown that a large number of patients with significant medical conditions were not referred for pre-anesthesia assessment. Furthermore, we evaluated the combined impact of information deficiency, patient complexity, and the surgical procedure complexity on providers' time in the APC. Based on these insights, we demonstrated that the APC could serve as PCSH system coordinator and handle the expected increase in demand with modest increases in resources. Since APC-like clinics are common in practice, our findings have great potential for widespread implementation of the PCSH model and significant benefits in terms of improved patient care and cost savings.

As part of ongoing research, we are doing a longitudinal study of various outcome measures as the PCSH is implemented. Preliminary data indicates that the number of APC monthly patients is increasing (in-clinic and telephone combined). In addition, the screening and triage is working, and the information quality is improving which has allowed APC to handle the increase in census as our simulation model predicted.

In future work, the PCSH model will be expanded further to include close coordination between anesthesiology and internal medicine. This collaboration would enable more effective optimization of patients' chronic medical conditions preoperatively, a safer intraoperative course, and improved postoperative care. Plans are in the works to move APC to a new facility that will house both anesthesiology and

internal medicine services, further strengthening the concept of a PCSH at UHS. Again, we will use a simulation model to forecast capacity and staffing levels in the expanded system.

Another area for future work involves improved patient scheduling. We intend to explore ways of scheduling patients in a coordinated fashion across multiple services providing another enhancement to the PCSH model. Finally, assessing the impact of other staffing models on the PCSH model could be an additional area of investigation.

ACKNOWLEDGMENTS

This work was supported by a grant from The University of Texas System. The authors would like to acknowledge and thank Ms. Carmen Sanchez and Mr. John Mark Atchley of UHS for providing data to support this project.

Chapter 3

Coordinated Patient Scheduling for a Multi-station Healthcare Network

3.1. Introduction

Annual spending on health in the United States is projected to grow 5.8% each year between 2014 and 2024. This growth rate is 1.1 times faster than the average GDP growth rate. It was estimated to hit $3.2 trillion dollars in 2015, which is already about 18% of the GDP (Bureau of Economic Analysis (2016); Centers for Medicare and Medicaid Services (2015)). Such spending is clearly unsustainable. Given the country's aging population, the outlook, needless to say, is troubling. Managing this increasing demand, when accompanied by tightening budget and resource scarcity, depends primarily on our ability to improve the efficiency and effectiveness of our care delivery. Recent healthcare reforms recognize this need and focus on moving toward patient-centered care to improve both efficiency and effectiveness (Centers for Medicare and Medicaid Services (2009)). All patient-centric models that have emerged advocate coordinating decisions among different services and providers, from resource planning to appointment scheduling.

Outpatient services account for more than four-fifths of patient care in the United States (Zeng et al. (2009)), and most patients access these services via appointment

scheduling. Hence, the need for efficient scheduling methods that coordinate multiple services to provide integrated episodes of patient care is abundantly clear. However, coordination of various outpatient services is difficult, both practically and technologically. Practically, challenges arise in modifying and training current providers that have traditionally operated as independent decision makers, managing their own schedules. Independent appointments for each service, separated on different dates, are currently the norm for patients. Even the coordination between various services within a hospital presents a challenge. Coordination between hospitals, then, is even more daunting. Technologically, we do not yet have the mathematical models, computational tools, and software implementations to allow such coordinated scheduling.

In a survey by Merritt Hawkins and Associates (2009), access delay for outpatient services ranges from weeks to months. Long delay in accessing services not only compromises the outcome of care but also leads to discontinuous care. Discontinuity in care leads to more expensive settings, like emergency rooms, where patients end up drawing more on the already strained resources. Another consequence of long-delayed access is the decrease in patient attendance rates, often referred to as patient no-shows. To compensate for patient no-shows and reduce provider idleness, services overbook their schedules (Chen and Robinson (2014)). Currently, as each service manages its own schedule, patients are left with the sequencing and the coordination of their various services over several weeks or months. This independence significantly amplifies no-shows because of cascading effects. Lack of coordination, among other inefficiencies, plays a central role in creating this down-

ward spiral: Service providers respond to increasing no-shows by overbooking clinics, which increases the cascading effect, as patients respond to long waits by higher chance of no-shows.

Such a spiral played a central role in the scandal at the Veteran Health Administration in 2014 that shocked the country. The VA was accused of systemic scheduling inefficiencies. A 2014 VA Inspector General's report (VA Report (2014)) concluded that deficiencies in scheduling was one of the leading causes of access delay that in some cases were more than 115 days. The report called for better scheduling policies that coordinate different services in the delivery of care. Another 2012 VA Inspector General's report (VA Report (2012)) emphasized coordination among the primary care doctor, nurse station and other staff in caring for patients as a way to improve efficiency.

An advantage of coordinated scheduling is the opportunity to anticipate and accommodate referrals before the patients' arrival. With this type of advanced planning, it is possible to schedule multiple services on a single patient visit, improving access to care and patient satisfaction. It also has the potential to reduce patient no-shows associated with referral appointments, mitigating uncertainty and thus reducing operational inefficiencies and costs in healthcare. Even healthcare reimbursement schemes are undergoing reform, moving from fee-for-service to bundled pricing for a single integrated episode of care that includes multiple services (Centers for Medicare and Medicaid Services (2009)). Coordinated scheduling works quite well with organizational changes that centralize appointment booking, enabled by improved health IT systems (Gupta and Denton (2008)).

In this chapter, we focus on filling the technological gap in accomplishing coordinated scheduling. Coordinated appointment scheduling is very limited in literature and to our knowledge, implementable models do not exist. As pointed out by Berg and Denton (2012), managing healthcare as a multi-station interconnected network is an open and important problem in the OR/Healthcare Management research. Our focus is on formulating a model and a computational methodology that can strike a balance between computational time and stylization. We formulate a stochastic network model that captures the complexity of patient no-shows, sequential scheduling necessity, service time uncertainty, and stochastic patient flows within and between services. Because stations in the network model represent services or clinics, we use these three terms interchangeably. A centralized scheduler uses patient information and preferences to make sequential appointment decisions that maximize a network objective. The objective is to balance the benefit of serving patients against the costs involved in patient waiting time and staff overtime.

Central to our methodology is the myopic scheduling approach. The myopic approach schedules each appointment request as if it is the last request received before the appointment day. The primary reason for adopting this approach is computational complexity. A dynamic programming formulation that relaxes the myopic policy is possible. However, the computational complexity due to the dimensionality of the state space, the types of decisions and the number of uncertainties, if at all possible, would result in astronomical computational times. Even if this task were achievable, any non-myopic methodology would require the specification of patient arrival processes and would yield a solution that is very sensitive to this

hard-to-calibrate arrival process. Moreover, unlike in the myopic case, a dynamic program requires a model for the slot preferences of future arriving patients. This is almost impossible to estimate with any available data, making the entire machinery futile. In general, a myopic decision would be close to optimal when nearing the end of the time horizon. However, in our situation, the myopic is also reasonable at initial stages as well, only because a very sparse schedule is forgiving to varying patient placements. In fact, we substantiate this argument by computing performance bounds and demonstrate that, for a range of parameters, the difference between our myopic approach and an unachievable super-optimal approach is within one percent.

In addition to the model formulation and the solution methodology, the chapter makes several other contributions. The most important among them is a sequence of approximation schemes. Even the myopic approach that has been proposed is time-consuming to compute and not reasonable to keep a patient waiting on the phone. Hence, for practical implementation using a reasonable desktop, faster methods are needed. We create a number of approximation schemes searching for ones that yield large computational advantages at very low approximation costs, and manage to find one such very beneficial approximation. The chapter also discusses and provides insights into the following: (1) how the complexity depends on the network structure; (2) how and why the different approximation schemes behave the way they do; and (3) the dependence of the network performance on various model parameters. For numerical illustrations and performance demonstrations, we use a simple but common network that captures all the necessary elements and manage-

rial insights we want to convey. For the sake of clearer insights and chapter length, we have refrained from an exhaustive set of computational examples on more complex networks, although the methodology and all approximations are readily applicable.

The rest of the chapter is structured as follows. Section 3.2 provides a brief review of the literature. In Section 3.3, we formulate the network model that captures the dynamics of patient flow within and between stations in the network. In Section 3.4, we derive the joint probability distributions for patient flow dynamics, which sheds light on the relation between problem complexity and network structure. Leveraging the analysis in Section 3.4, Section 3.5 lays out the myopic approach to solving the scheduling problem. We also discusses various important implementation aspects. Approximation schemes designed to overcome the computational challenges are discussed in Section 3.6. Section 3.7 aggregates the computational studies that illustrate the scheduling algorithm, discusses the quality-efficiency trade-off among different approximation schemes, provides the optimality gaps and discusses other insights. We make our concluding remarks in Section 3.8.

3.2. Related Literature

Literature on healthcare scheduling predominantly studies single-station environments. In this section, we begin by providing an objective and concise overview of the literature on single-station scheduling. We group and discuss papers based on their modeling choices and assumptions. Cayirli and Veral (2003), Gupta and

Denton (2008) and Gupta and Wang (2012) provide excellent and comprehensive reviews of much of this literature on single stations. However, we see far less literature on scheduling in multi-station settings and almost all the available literature focuses on evaluating heuristic policies using simulation or on optimizing parameters of heuristic policies using simulation optimization. To conclude this section, we discuss all the papers we could find on this topic in the current literature.

Probably the most important modeling choice is between on-line and off-line scheduling, also referred to as sequential and static scheduling, respectively. Off-line scheduling assumes that a fixed number of patients (usually homogenous with no preferences) are to be scheduled. In sequential scheduling, as in practice, patients call in sequentially before the day of service to make an appointment. Decisions are made one at a time in the order of request arrivals, with limited or no information about future requests. Hence, the total number of patients to be scheduled is uncertain but partly controllable because patients can be refused an appointment for a particular day. Although more realistic, sequential scheduling does not offer the same level of analytical tractability offered by static scheduling.

Statics models have enjoyed significant attention in the literature. Given the number of surgeries to be scheduled, Denton and Gupta (2003) study the optimal sequencing of surgeries based on their estimated durations, using a stochastic linear programming model. % is formulated to allocate block-times for surgeries that optimizes resource utilization and minimizes delays. Erdogan and Denton (2013) extend Denton and Gupta (2003) by allowing patient no-shows. Kaandorp and Koole (2007) propose an algorithm that locally improves any given schedule and show that the

solution is optimal when objective functions are multi-modular. Hassin and Mendel (2008) develop a single-server queueing model to optimize the appointment time for a fixed number of patients and obtain a closed-form solution for a two-patient system. LaGanga and Lawrence (2012) propose a gradient search heuristic and show numerically that its solution is near-optimal.

In sequential scheduling, a closed-form analytical solution is almost never possible. Due to the challenges in dimensionality, even computational methods for optimization become intractable. Hence, researchers primarily have relied on performance evaluation methods, like simulation or heuristics. Klassen and Rohleder (1996) use simulation to find scheduling rules that reserve appointment slots for urgent patients, as opposed to non-urgent patients. Klassen and Yoogalingam (2009) consider more factors in the decision-making process and propose a schedule that they call a robust plateau-dome scheduling pattern. Erdogan and Denton (2013) develop a multi-stage linear program that dynamically assigns patient appointment on a FCFS basis. In a later paper, Erdogan et al. (2015) consider the problem of sequencing patient appointments where requests are not FCFS and the scheduler reserves capacity in anticipation of urgent requests with short notice. Zacharias and Pinedo (2014) study the structure of optimal schedules under heterogeneous patient no-show types. They first derive the optimal schedule for the static model, from which they gain insights to design a heuristic for the sequential model. Feldman et al. (2014) adopt a similar strategy. In addition, they define an optimality bound to evaluate their proposed heuristic. The single-station model considered in Muthuraman and Lawley (2008) uses the same slot structure that we use for each clinic

in this chapter. They show that a simple myopic policy lends itself well to a stopping criterion that provides the convenience of theoretical characterization. This is largely obtained by first establishing that their objective is unimodal. In a multi-station setting, such unimodality does not exist, making almost all of the analysis in Muthuraman and Lawley (2008) futile for our purposes.

Patient no-shows greatly affect operational efficiencies, causing loss of provider productivity. Overbooking hedges against the risk of resource idleness, but it can cause longer waiting times. And together, no-shows and overbooking often result in long access delays. Using simulation, LaGanga and Lawrence (2007) suggest that the amount of overbooked appointments should increase as no-show probability increases. In a later paper, LaGanga and Lawrence (2012) formulate an optimization problem to find the optimal amount of overbooking to balance patient no-shows.

In addition to patient no-shows, unexpected patient arrivals (referred to as walk-ins and add-ons) are another factor that disrupts clinical operations. If not properly accounted for, these disturbances can cause significant operational inefficiencies and patient delays (Luo et al. (2012) and Cayirli et al. (2012)). Cayirli et al. (2012) model the occurrence of add-on patients by its mean and standard deviation and develop a dome-shaped appointment rule that adjusts appointment durations based on the walk-in rate and no-show realizations. Luo et al. (2012) model the urgent requests that randomly arrive throughout the day as a time-dependent Poisson process. They formulate an optimization problem and use simulation optimization. Chen and Robinson (2014) formulate the problem as a stochastic linear program and obtain appointment sequences for small problem instances.

Another significant modeling assumption is the choice of deterministic or stochastic service times. Even though deterministic service times can be argued to be unrealistic, this assumption allows significant tractability. By assuming that a patient's service time is identical and equal to the duration of an appointment, Zacharias and Pinedo (2014) derive an optimal schedule for homogeneous patient service times. They then propose a heuristic for heterogeneous service times based on the insights from the homogeneous model. For similar reasons, other works that assume deterministic service times include Robinson and Chen (2010), LaGanga and Lawrence (2007) and LaGanga and Lawrence (2012). In terms of stochastic service times, a popular choice is the exponential distribution, because of its analytical tractability and its capability of modeling high variation in service times, which is commonly observed in healthcare systems. For example, using exponential service time, Wang (1993) is able to show the optimality of a dome-shape scheduling rule. Exponential service times, being the only memoryless distribution, do offer a great theoretical and computational advantage. In this chapter we stick with the exponential distribution primarily because of its computational advantages and its ability to provide valuable insights. Chakraborty et al. (2010) consider a general distribution and demonstrate the computational complexity involved. They also compare the log-normal, gamma and exponential distributions and evaluate them. Klassen and Yoogalingam (2009) and Zacharias and Pinedo (2014) also consider log-normal service times but due to limited analytical tractability of the log-normal distribution, they rely on simulation for their analysis.

Finally, we move on to a relatively sparse multi-station setting literature. White

et al. (2011) model an outpatient clinic that consists of a sequence of services. They use simulation to study how policies on room allocation and appointment scheduling jointly affect clinic performance. Unlike our model, appointments requests in White et al. (2011) are for the clinic, not for a particular time slot, and patient routing is predetermined. Chao et al. (2003) develop an analytical model that allots patients to substitutable sites in a multi-site healthcare system, but the model does not include scheduling, and patients do not visit one station after another. In the supply chain scheduling literature, a series of papers by Hall and Potts (2003); Chen and Hall (2007); Dawande et al. (2009) consider coordinated scheduling in supply chains. However, these papers assume deterministic models and may be viewed as extensions of the two-stage flow shop models to coordinated scheduling decisions among different players in the supply chain.

3.3. Model Formulation

Consider a network consisting of I stations with schedules that are equally divided into J time slots for appointment booking. A station could be a service center (e.g., lab, X-ray, or physician's office) or a clinic in a co-located medical district that interacts through inter-station patient referrals. We use (i, j) to denote station i slot j and use **I** and **J** to denote the set of stations and appointment slots in the network, respectively. Patients who need an appointment call a centralized scheduler prior to the beginning of the appointment period. In practice, an appointment period is typically a day, but it can also be a morning or afternoon session within a day. For convenience, we refer to an appointment period as a day. Appointment slots are, for

example, 30-minute divisions of an appointment period. Appointment periods are independent because all services are completed within each period and healthcare providers work overtime, if necessary. As scheduling patients on different days is equivalent to solving identical problems in parallel, we focus on requests for one day.

When a call arrives, the centralized scheduler asks for the type of service/station ($i \in \mathbf{I}$) the patient needs and the preferred time slots, defined as set $\mathbf{C}$. Either the patient is scheduled into one of the preferred slots or the request is not accepted for that day and is transferred to another appointment day. A network schedule, $\mathbf{S}$, is an $I \times J$ matrix that specifies the number of patients scheduled in each slot-station pair. Scheduled patients might or might not show up for an appointment. Hence, the actual number of exogenous arrivals at (i, j), denoted as $X_{i,j}$, is a random variable $X_{i,j} \sim \text{Binomial}(S_{i,j}, \rho_i)$, where ρ_i is the station-dependent patient attendance probability. For notational convenience, we leave the patient no-show rates as station-dependent only. A patient-specific no-show probability can be incorporated by including a patient-type dimension to S, with each type having a no-show probability. Service times are assumed to be exponentially distributed, with station-dependent service rate λ_i. At the end of each slot, a patient either completes service or overflows to the next slot. Let $Z_{i,j}$ be the number of completed services from (i, j), and let $Y_{i,j}$ be the number of patients who overflow to $(i, j+1)$.

After the service is complete, the patient either gets routed to another station or exits the network, according to a probabilistic routing matrix, $\mathbf{P}$. Let $P_{i,k}$ be the probability a patient is referred to station k when the patient completes service in station

i. Inter-station patient flows are called referrals. The patient leaves the network after station i with probability $1 - \sum_{k \neq i} P_{i,k}$. No immediate re-entrance is allowed in the network. The assumption of stochastic routing allows for analytical tractability by obviating the need to track individual patient paths. It also reflects the uncertainty in a patient's health condition prior to each medical visit. As a result of coordinated scheduling, we assume referral patients can join the queue at the corresponding service in the immediate subsequent slot.

$$\mathbf{R}_{\cdot,\cdot,\mathbf{j}} \equiv \begin{bmatrix} R_{1,1,j} & R_{1,2,j} & ... & R_{1,I,j} \\ R_{2,1,j} & R_{2,2,j} & ... & R_{2,I,j} \\ \vdots & & \ddots & \vdots \\ R_{I,1,j} & R_{I,2,j} & ... & R_{I,I,j} \end{bmatrix},$$

where the random variable $R_{i,k,j}$ is the number of referrals from (i,j) to $(k,j+1)$. Given the number of services completed at (i,j), we have the following multinomial distribution

$$(\mathbf{R}_{\mathbf{i},\cdot,\mathbf{j}} \quad Z_{i,j} - \sum_{k \in \mathbf{I}} R_{i,k,j}) | Z_{i,j} \sim \text{multinomial}(Z_{i,j}, P_{i,1}, P_{i,2}, ..., P_{i,I}, 1 - \sum_{k \in \mathbf{I}} P_{i,k}), \tag{3.1}$$

where the last term is the probability that patient's exiting the network. We use the vector form and the dot notation to represent an array of variables that share one identical subscript.

Figure 3.1 shows the different types of patient flows in and out of (i,j) and the interdependence among slots and stations. Note that the number of patients at (i,j) depends on the number of patients leaving from other stations and being referred to (i,j), the number of overflows from the previous slot, and the number of exogenous arrivals. We use $A_{i,j}$ to denote the total number of patients arriving at (i,j). We have

$$X_{i,j} + Y_{i,j-1} + \sum_{k \in \mathbf{I}} R_{k,i,j-1} = A_{i,j} = Z_{i,j} + Y_{i,j} \quad \forall i \in \mathbf{I},\, j \in \mathbf{J}\,, \tag{3.2}$$

where $Y_{i,0} = 0$. Let L_i be the number of services that can be completed in a slot at station i, given a non-empty queue. Then the number of services completed in (i,j) is

$$Z_{i,j} = \min(L_i, A_{i,j})\,, \tag{3.3}$$

and the number of overflows from (i,j) to $(i,j+1)$ is

$$Y_{i,j} = \max(0,\, A_{i,j} - L_i). \tag{3.4}$$

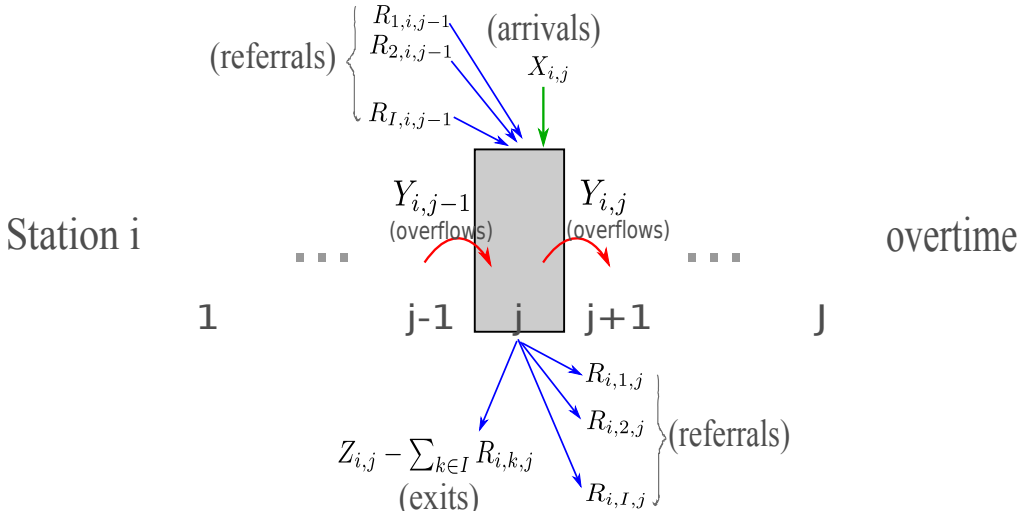

Figure 3.1: Patient flow dynamics

The objective of the centralized scheduler is to strike a good balance between wait times and provider utilization. In making each appointment decision, the scheduler needs to consider the cost implications of the patient's not showing up

for the appointment and the implications of possible congestion. Moreover, the decision has to be made based on limited information about future requests, patient preferences and no-show probabilities. As can be seen, both the dimensionality of the state space and the number of sources of randomness rule out the use of standard dynamic programming.

As discussed in the introduction, we propose a myopic policy since dynamic programs are too complex and sensitive. Moreover, myopic policies tend to perform very well in this context. For a range of reasonable parameters, in Section 3.7, we demonstrate that the difference between our myopic approach and an unachievable super-optimal approach is within one percent.

The objective under a myopic policy is to assign each patient to a slot so that the resulting schedule maximizes a profit function when no future requests are considered. The profit function for a schedule is denoted as $V_T(\mathbf{S})$, which is expressed as the difference between the expected revenue from all services provided by the network and the expected cost of patient waiting and staff overtime, i.e. $V_T(\mathbf{S}) = \mathrm{E}(\text{Revenue}(\mathbf{S})) - \mathrm{E}(\text{Cost}(\mathbf{S}))$. The expected revenue for a particular $\mathbf{S}$ is the sum of rewards for providing nominal and referral services, accounting for patient no-show, and is expressed as

$$\mathrm{E}(\text{Revenue}(\mathbf{S})) = \sum_{i \in \mathbf{I}} \mathbf{r}[\mathbf{I} - \mathbf{P}']^{-1} \hat{e}_i \left(\sum_j S_{i,j} \rho_i \right), \tag{3.5}$$

where $\mathbf{r} = (r_1 r_2 ... r_I)$ and r_i is the reward at station i. In a for-profit setting, a reward can be viewed as the service fee; while in a non-profit setting, a reward can be seen as the utility for serving a patient (LaGanga and Lawrence (2007, 2012)). The inverse

matrix $[\mathbf{I} - \mathbf{P}']^{-1}$ computes the expected number of visits to each station a patient demands before exiting the network. The notation $\hat{e}_i$ is a $I \times 1$ unit vector whose ith row is of value 1. The effect of $\hat{e}_i$ is to extract the ith column of $[\mathbf{I} - \mathbf{P}']^{-1}$ for vector manipulation.

As is standard in multi-objective models, we penalize the different objectives with an exogenously specified coefficient that captures the relative reference between the various objectives. Penalizing the objective by a cost of $c_{i,j}$ when a patient overflows from slot j ($j < J$) to $j+1$, the expected cost of a schedule becomes

$$\mathrm{E}\left(\mathrm{Cost}(\mathbf{S})\right) \;=\; \sum_{j=1}^{J} \sum_{\mathbf{y}_{.,\mathbf{j}}} P(\mathbf{Y}_{.,\mathbf{j}} = \mathbf{y}_{.,\mathbf{j}}) \mathbf{y}_{.,\mathbf{j}} \mathbf{c}'_{.,\mathbf{j}}\,, \tag{3.6}$$

where $\mathbf{Y}_{.,\mathbf{j}} \equiv (Y_{1,j} = y_{1,j},\, Y_{2,j} = y_{2,j}, ...,\, Y_{I,j} = y_{I,j})$ and $\mathbf{c}_{.,\mathbf{j}} = (c_{1,j}\; c_{2,j} ..., c_{I,j})$. Here, $\mathbf{c}_{.,\mathbf{J}}$ represents the penalty when a patient overflows to clinic overtime. The notation $\sum_{\mathbf{y}_{.,\mathbf{j}}}$ represents a summation over all realizations of the joint overflow variables at slot j.

3.4. Joint Overflow Distribution

To evaluate the objective for any given schedule, we need the joint distribution of overflow variables. Figure 3.2 reveals the recursive structure involved in deriving the joint-overflow distribution, given any schedule. It illustrates the complexity in evaluating $P(\mathbf{Y}_{.,\mathbf{j}})$. In this section we first derive the unconditional overflow distribution of $\mathbf{Y}_{.,j}$ and express it in terms of the distribution of $\mathbf{Z}_{.,\mathbf{j}}$ and the conditional distribution of $\mathbf{Y}_{.,j}|\mathbf{Z}_{.,j}$, which are presented in Propositions 3.4.2 and 3.4.3, respectively. Together, these distributions establish the recursive dependence structure for

the computation. For notation simplicity, we omit the conditional term, $|\mathbf{S}$, from all the expressions in the following paragraphs. For example, $P(\mathbf{Y}_{\cdot,\mathbf{j}}|\mathbf{S})$ is abbreviated as $P(\mathbf{Y}_{\cdot,\mathbf{j}})$. As shown in Figure 3.1, a single overflow variable at (i, j), $Y_{i,j}$, is related to overflows from the previous slot, $Y_{i,j-1}$; referrals that are due to arrive, $\mathbf{R}_{\cdot,i,j-1}$; and the exogenous arrivals, $X_{i,j}$. Figure 3.2 serves as a roadmap for the following derivations of the distribution of $P(\mathbf{Y}_{\cdot,j})$.

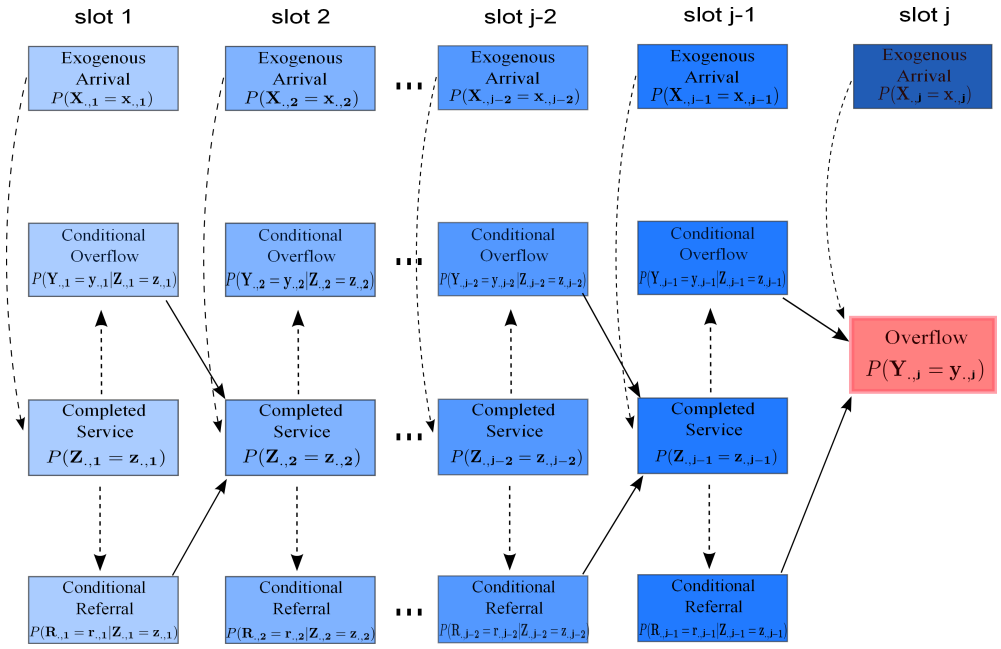

Figure 3.2: Recursive Structure of the Joint Overflow Distribution

Proposition 3.4.1. *The joint overflow distribution,* $P(\mathbf{Y}_{\cdot,j} = \mathbf{y}_{\cdot,j})$*, is*

$$P(\mathbf{Y}_{\cdot,1} = \mathbf{y}_{\cdot,1}) = \prod_{i \in \mathbf{I}} \sum_{x_{i,1}=0}^{S_{i,1}} P(X_{i,1} = x_{i,1}) \cdot \left\{ \begin{array}{c} 1(y_{i,1} = 0) P(L_i \geqslant x_{i,1}) \\ + \\ 1(y_{i,1} > 0) P(L_i = x_{i,1} - y_{i,1}) \end{array} \right\},$$

for $j = 1$, and for $j > 1$, it is

$$
\begin{aligned}
P(\mathbf{Y}_{\cdot,\mathbf{j}} = \mathbf{y}_{\cdot,\mathbf{j}}) \;=\; & \sum_{\mathbf{a}_{\cdot,\mathbf{j}} \in \Omega_a^j} \prod_{i \in \mathbf{I}} \left\{ \begin{array}{c} 1(y_{i,j} > 0) \cdot P(L_i = a_{i,j} - y_{i,j}) \\ + \\ 1(y_{i,j} = 0) P(L_i \geq a_{i,j} - y_{i,j}) \end{array} \right\} \cdot \\
& \prod_{i \in \mathbf{I}} \sum_{x_{i,j}=0}^{S_{i,j}} P(X_{i,j} = x_{i,j}) \Bigg(\sum_{\mathbf{z}_{\cdot,\mathbf{j-1}} \in \mathbf{\Omega_z^{j-1}}} P(\mathbf{Z}_{\cdot,\mathbf{j-1}} = \mathbf{z}_{\cdot,\mathbf{j-1}}) \cdot \\
& \Big(\sum_{\mathbf{r}_{\cdot,\cdot,\mathbf{j-1}} \in \mathbf{\Omega}_{\mathbf{r}|\mathbf{z}}^{\mathbf{j-1}}} P(\mathbf{R}_{\cdot,\cdot,\mathbf{j-1}} = \mathbf{r}_{\cdot,\cdot,\mathbf{j-1}} | \mathbf{z}_{\cdot,\mathbf{j-1}}) \cdot \\
& P(\mathbf{Y}_{\cdot,\mathbf{j-1}} = \mathbf{a}_{\cdot,\mathbf{j}} - \mathbf{x}_{\cdot,\mathbf{j}} - \mathbf{r}'_{\cdot,\cdot,\mathbf{j-1}} \cdot \mathbf{I}_{I \times 1} \mid \mathbf{z}_{\cdot,\mathbf{j-1}}) \Big) \Bigg).
\end{aligned}
$$

Proof. When $j = 1$, by the balance of flow, we have $\mathbf{Y}_{\cdot,\mathbf{1}} = \mathbf{X}_{\cdot,\mathbf{1}} - \mathbf{Z}_{\cdot,\mathbf{1}}$. Conditioning on the number of patients that show up, we have

$$
\begin{aligned}
P(\mathbf{Y}_{\cdot,1} \;=\; \mathbf{y}_{\cdot,1}) &= \sum_{\mathbf{x}_{\cdot,1} \in \mathbf{S}_{\cdot,\mathbf{1}}} P(\mathbf{X}_{\cdot,1} = \mathbf{x}_{\cdot,1}) P(\mathbf{Y}_{\cdot,1} = \mathbf{y}_{\cdot,1} | \mathbf{X}_{\cdot,1} = \mathbf{x}_{\cdot,1}) \\
&= \prod_{i \in \mathbf{I}} \sum_{x_{i,1}=0}^{S_{i,1}} P(X_{i,1} = x_{i,1}) P(Y_{i,1} = y_{i,1} | X_{i,1} = x_{i,1}) \\
&= \prod_{i \in \mathbf{I}} \sum_{x_{i,1}=0}^{S_{i,1}} P(X_{i,1} = x_{i,1}) P(max(x_{i,j} - L_i, 0) = y_{i,1} | X_{i,1} = x_{i,1}) \\
&= \prod_{i \in \mathbf{I}} \sum_{x_{i,1}=0}^{S_{i,1}} P(X_{i,1} = x_{i,1}) \cdot \left\{ \begin{array}{c} 1(y_{i,1} = 0) P(L_i \geqslant x_{i,1}) \\ + \\ 1(y_{i,1} > 0) P(L_i = x_{i,1} - y_{i,1}) \end{array} \right\}.
\end{aligned}
\tag{3.7}
$$

The last equality follows from Equation (3.2) and from $L_i \sim Poisson(\lambda_i)$. The set of indicator functions represent the following: if the provider has sufficient capacity, all patients will be served and only the first indicator function returns a value 1. Otherwise, overflows occur and only the second indicator function becomes 1.

For $j > 1$, we first condition on the number of patients present in each station. By the balance of flow, $\mathbf{A}_{.,j} = \mathbf{Y}_{.,j} + \mathbf{Z}_{.,j}$, we have

$$\begin{aligned} P(\mathbf{Y}_{.,j} &= \mathbf{y}_{.,j}) = \sum_{\mathbf{a}_{.,j} \in \Omega_a^j} P(\mathbf{A}_{.,j} = \mathbf{a}_{.,j}) P(\mathbf{Y}_{.,j} = \mathbf{y}_{.,j} | \mathbf{A}_{.,j} = \mathbf{a}_{.,j}) \\ &= \sum_{\mathbf{a}_{.,j} \in \Omega_a^j} P(\mathbf{A}_{.,j} = \mathbf{a}_{.,j}) P(\mathbf{Z}_{.,j} = \mathbf{a}_{.,j} - \mathbf{y}_{.,j} | \mathbf{A}_{.,j} = \mathbf{a}_{.,j}) \\ &= \sum_{\mathbf{a}_{.,j} \in \Omega_a^j} P(\mathbf{A}_{.,j} = \mathbf{a}_{.,j}) \prod_{i \in \mathbf{I}} P(Z_{i,j} = a_{i,j} - y_{i,j} | A_{i,j} = a_{i,j}). \end{aligned} \tag{3.8}$$

The notation $\sum_{\mathbf{a}_{.,j}}$ represents the summation over the combinations of all possible values of $a_{i.,j}$ $\forall i \in \mathbf{I}$ and Ω_a^j is the support set for $\mathbf{A}_{.,\mathbf{j}}$. To define this support, we need to first define the upper bound on the number of patients that can reach $(i,\ j)$, denoted as $U(i,\ j)$. It is the sum of the scheduled patients in previous slots across all stations, plus the ones scheduled for $(i,\ j)$:

$$U(i,j) \equiv \sum_{i \in \mathbf{I}} \sum_{j' < j} S_{i,j'} + S_{i,j}. \tag{3.9}$$

The support set for $\mathbf{A}_{.,\mathbf{j}}$ is defined as:

$$\Omega_a^j \coloneqq \left\{ \times_{i \in \mathbf{I}} \left\{ a_{i,j} \in \mathbb{Z} :\ 0 \leq a_{i,j} \leq U(i,\ j) \right\} \right\} \bigcap \left\{ \sum_{i \in \mathbf{I}} a_{i,j} \leq \sum_{i \in \mathbf{I}} \sum_{j' \leq j} S_{i,j'} \right\}.$$

The first set is a cartesian product of the range of values for each $a_{i.,j}$. The second set ensures that the joint distribution is feasible. i.e., the total number of patients in the network at slot j is no greater than the total number of patients scheduled up to j. The last line in Equation (3.8) follows from the independence of $Z_{i,j}$ when conditioned on the $A_{i,j}$. Next, we derive $P(\mathbf{A}_{.,j} = \mathbf{a}_{.,j})$ and $P(Z_{i,j} = z_{i,j} | A_{i,j} = a_{i,j})$.

By Equation (3.2), we have $P(\mathbf{A}_{\cdot,1} = \mathbf{a}_{\cdot,1}) = P(\mathbf{X}_{\cdot,1} = \mathbf{a}_{\cdot,1}) = \prod_{i\in\mathbf{I}} P(X_{i,1} = a_{i,1})$, for $j = 1$. For $j > 1$, conditioning on the exogenous arrivals, we have

$$\begin{aligned}
P(\mathbf{A}_{\cdot,j} &= \mathbf{a}_{\cdot,j}) = P(\mathbf{X}_{\cdot,j} + \mathbf{Y}_{\cdot,j-1} + \mathbf{R}_{\cdot,\cdot,j-1} = \mathbf{a}_{\cdot,j}) \\
&= \prod_{i\in\mathbf{I}} \sum_{x_{i\cdot,j}=0}^{S_{i,j}} P(X_{i.j} = x_{i,j}) P(\mathbf{Y}_{\cdot,j-1} + \mathbf{R}_{\cdot,\cdot,j-1} = \mathbf{a}_{\cdot,j} - \mathbf{x}_{\cdot,j} | \mathbf{X}_{\cdot,j} = \mathbf{x}_{\cdot,j}) \\
&= \prod_{i\in\mathbf{I}} \sum_{x_{i\cdot,j}=0}^{S_{i,j}} P(X_{i.j} = x_{i,j}) \sum_{\mathbf{z}_{\cdot,j-2}\in\Omega_z^{j-2}} P(\mathbf{Z}_{\cdot,j-1} = \mathbf{z}_{\cdot,j-1}) \\
&\quad P(\mathbf{Y}_{\cdot,j-1} + \mathbf{R}_{\cdot,\cdot,j-1} = \mathbf{a}_{\cdot,j} - \mathbf{x}_{\cdot,j} | \mathbf{z}_{\cdot,j}) \\
&= \prod_{i\in\mathbf{I}} \sum_{x_{i\cdot,j}=0}^{S_{i,j}} P(X_{i.j} = x_{i,j}) \sum_{\mathbf{z}_{\cdot,j-1}\in\Omega_z^{j-1}} P(\mathbf{Z}_{\cdot,j-1} = \mathbf{z}_{\cdot,j-1}) \\
&\quad \sum_{\mathbf{r}_{\cdot,\cdot,j-1}\in\Omega_r^{j-1}} P(\mathbf{R}_{\cdot,\cdot,j-1} = \mathbf{r}_{\cdot,\cdot,j-1} | \mathbf{z}_{\cdot,j-1}) \\
&\quad P(\mathbf{Y}_{\cdot,j-1} = \mathbf{a}_{\cdot,j} - \mathbf{x}_{\cdot,j} - \mathbf{r}'_{\cdot,\cdot,j-1} \cdot I_{I\times 1} | \mathbf{z}_{\cdot,j-1}).
\end{aligned} \tag{3.10}$$

The third equality in Equation (3.10) results from conditioning $(\mathbf{Y}_{\cdot,j-1}, \mathbf{R}_{\cdot,\cdot,j-1})$ on $\mathbf{Z}_{\cdot,j-1}$. By independence, the conditional term on $\mathbf{X}_{\cdot,j}$ can be dropped. Furthermore, given $Z_{i,j-1}$ the distribution of $\mathbf{R}_{i,\cdot,j-1}$ is independent of $\mathbf{Y}_{\cdot,j-1}$, which results the last equality in Equation (3.10). The notation $\sum_{\mathbf{z}_{\cdot,\mathbf{j-1}}}$ represents the summation over the combinations of all possible values of $z_{i\cdot,j-1}$, $\forall i \in \mathbf{I}$. The support set for $\mathbf{Z}_{\cdot,\mathbf{j-1}}$ is Ω_z^{j-1}, which is defined as the following:

$$\Omega_z^{j-1} := \left\{ \times_{i\in\mathbf{I}} \left\{ z_{i,j-1} \in \mathbb{Z} : \, 0 \le z_{i,j-1} \le \sum_{i\in\mathbf{I}} U_{i,j-1} \right\} \right\} \bigcap \left\{ \sum_{i\in\mathbf{I}} z_{i,j-1} \le \sum_{i\in\mathbf{I}} \sum_{j'\le j-1} S_{i,j'} \right\}.$$

Similar to Ω_a^j, the first set is a cartesian product of the range of values for each station. The second set ensures that the joint distribution is feasible - the total number of completed services at slot $j-1$ is no greater than the total number of patients scheduled up to slot $j-1$.

For $P(Z_{i,j} = z_{i,j}|A_{i,j} = a_{i,j})$, Equation (3.3) yields the following:

$$\begin{aligned} P(Z_{i,j} &= z_{i,j}|A_{i,j} = a_{i,j}) = P(Y_{i.,j} = a_{i,j} - z_{i,j}|A_{i,j} = a_{i,j}) \\ &= P(max(0,\, a_{i,j-1} - L_i) = a_{i,j} - z_{i,j-1}) \\ &= \prod_{i \in \mathbf{I}} \left\{ \begin{array}{c} 1(a_{i,j-1} - z_{i,j-1} = 0) P(L_i \geqslant z_{i,j-1}) \\ + \\ 1(a_{i,j-1} - z_{i,j-1} > 0) P(L_i = z_{i,j-1}) \end{array} \right\}. \end{aligned} \tag{3.11}$$

The set of indicator functions can be interpreted similarly to that in Equation (3.7).

To conclude the proof, substituting Equations (3.10) and (3.11) into Equation (3.8) yields Proposition 3.4.1 for $j > 1$. □

Proposition 3.4.1 provides insight into the recursive structure and the complexity of the algorithm. To obtain the joint overflow distribution at any slot $j > 1$, we need to know the probability distributions of the exogenous arrivals, conditional referrals, completed services and conditional overflows. The $X_{i,j} \sim Binomial(S_{i,j},\, \rho_i)$ and $\mathbf{R}_{.,.,j}|\mathbf{Z}_{.,j}$ can be obtained from a multinomial distribution (Equation (3.1)). The other two probability distributions, $P(\mathbf{Z}_{.,j} = \mathbf{z}_{.,j})$ and $P(\mathbf{Y}_{.,j} = \mathbf{y}_{.,j}|\mathbf{Z}_{.,j} = \mathbf{z}_{.,j})$, are derived in Propositions 3.4.2 and 3.4.3, respectively.

Proposition 3.4.2. *The joint service completion distribution, $P(\mathbf{Z}_{.,j} = \mathbf{z}_{.,j})$, is*

$$P(\mathbf{Z}_{.,1} = \mathbf{z}_{.,1}) = \prod_{i \in \mathbf{I}} \sum_{x_{i,1}=0}^{S_{i,1}} P(X_{i,1} = x_{i,1}) \left\{ \begin{array}{c} 1(z_{i,1} = x_{i,1}) P(L_i \geqslant z_{i,1}) \\ + \\ 1(z_{i,1} < x_{i,1}) P(L_i = z_{i,1}) \end{array} \right\},$$

for $j = 1$, and for $j > 1$, it is

$$
\begin{aligned}
P(\mathbf{Z}_{\cdot,\mathbf{j}} = \mathbf{z}_{\cdot,\mathbf{j}}) \;=\; & \sum_{\mathbf{a}_{\cdot,\mathbf{j}} \in \mathbf{\Omega_a^j}} \prod_{i \in \mathbf{I}} \left\{ \begin{array}{c} 1(z_{i,j} = a_{i,j}) \cdot P(L_i \geq z_{i,j}) \\ + \\ 1(z_{i,j} < z_{i,j}) \cdot P(L_i = z_{i,j}) \end{array} \right\} \cdot \\
& \prod_{i \in \mathbf{I}} \sum_{x_{i,j}=0}^{S_{i,j}} P(X_{i,j} = x_{i,j}) \Bigg(\sum_{\mathbf{z}_{\cdot,\mathbf{j-1}} \in \mathbf{\Omega_z^{j-1}}} P(\mathbf{Z}_{\cdot,\mathbf{j-1}} = \mathbf{z}_{\cdot,\mathbf{j-1}}) \cdot \\
& \Big(\sum_{\mathbf{r}_{\cdot,\cdot,\mathbf{j-1}} \in \mathbf{\Omega}_{\mathbf{r|z}}^{\mathbf{j-1}}} P(\mathbf{R}_{\cdot,\cdot,\mathbf{j-1}} = \mathbf{r}_{\cdot,\cdot,\mathbf{j-1}} | \mathbf{z}_{\cdot,\mathbf{j-1}}) \cdot \\
& \left(P(\mathbf{Y}_{\cdot,\mathbf{j-1}} = \mathbf{a}_{\cdot,\mathbf{j}} - \mathbf{x}_{\cdot,\mathbf{j}} - \mathbf{r}'_{\cdot,\cdot,\mathbf{j-1}} \cdot \mathbf{I}_{I \times 1} \mid \mathbf{z}_{\cdot,\mathbf{j-1}}) \right) \Big) \Bigg).
\end{aligned}
$$

Proof. For $j = 1$, the completed service variables are independent across stations in the first slot, and we have

$$
\begin{aligned}
P(\mathbf{Z}_{\cdot,1} \;=\; \mathbf{z}_{\cdot,1}) &= \prod_{i \in \mathbf{I}} P(Z_{i,1} = z_{i,1}) \\
&= \prod_{i \in \mathbf{I}} \sum_{x_{i,1}=0}^{S_{i,1}} P(X_{i,1} = x_{i,1}) \cdot P(min(L_i,\, x_{i,j}) = z_{i,1} | X_{i,1} = x_{i,1}) \\
&= \prod_{i \in \mathbf{I}} \sum_{x_{i,1}=0}^{S_{i,1}} P(X_{i,1} = x_{i,1}) \cdot \left\{ \begin{array}{c} 1(z_{i,1} = x_{i,1}) P(L_i \geqslant z_{i,1}) \\ + \\ 1(z_{i,1} < x_{i,1}) P(L_i = z_{i,1}) \end{array} \right\}. \qquad (3.12)
\end{aligned}
$$

The second line holds by Equation (3.3). For $j > 1$, we first condition on the total number of patients at slot j. By the independence of $Z_{i,j}|A_{i,j}$ across stations and Equation (3.11), we have:

$$
\begin{aligned}
P(\mathbf{Z}_{\cdot,j} \;=\; \mathbf{z}_{\cdot,j}) &= \sum_{\mathbf{a}_{\cdot,\mathbf{j}} \in \mathbf{\Omega_a^j}} P(\mathbf{A}_{\cdot,j} = \mathbf{a}_{\cdot,j}) \prod_{i \in \mathbf{I}} P(Z_{i,j} = z_{i,j} | A_{i,j} = a_{i,j}) \\
&= \sum_{\mathbf{a}_{\cdot,\mathbf{j-1}} \in \mathbf{\Omega_a^j}} P(\mathbf{A}_{\cdot,j} = \mathbf{a}_{\cdot,j}) \prod_{i \in \mathbf{I}} \left\{ \begin{array}{c} 1(z_{i,j} = a_{i,j}) P(L_i \geqslant z_{i,j}) \\ + \\ 1(z_{i,j} < a_{i,j}) P(L_i = z_{i,j}) \end{array} \right\}. \qquad (3.13)
\end{aligned}
$$

Substituting Equation (3.10) into Equation (3.13) concludes the proof. □

Proposition 3.4.3. *The joint conditional overflow distribution,* $P(\mathbf{Y}_{.,j} = \mathbf{y}_{.,j} | \mathbf{Z}_{.,j} = \mathbf{z}_{.,j})$*, is*

$$P(\mathbf{Y}_{.,1} = \mathbf{y}_{.,1} | \mathbf{Z}_{.,1} = \mathbf{z}_{.,1}) = \prod_{i \in \mathbf{I}} \left\{ \begin{array}{c} 1(y_{i,1} = 0) P(L_i \geqslant z_{i,j}) \\ + \\ 1(y_{i,1} > 0) P(L_i = z_{i,j}) \end{array} \right\},$$

for $j = 1$, and for $j > 1$, it is

$$\begin{aligned} P(\mathbf{Y}_{.,\mathbf{j}} = \mathbf{y}_{.,\mathbf{j}} | \, \mathbf{Z}_{.,\mathbf{j}} = \mathbf{z}_{.,\mathbf{j}}) \;\; = \;\; & \frac{1}{\sum_{\mathbf{z'}_{.,\mathbf{j}} \mathbf{\Omega}_{\mathbf{z}}^{\mathbf{j}}} P(\mathbf{Z}_{.,\mathbf{j}} = \mathbf{z'}_{.,\mathbf{j}})} \prod_{i \in \mathbf{I}} \left\{ \begin{array}{c} 1(y_{i,j} = 0) \cdot P(L_i \geq z_{i,j}) \\ + \\ 1(y_{i,j} > 0) \cdot P(L_i = z_{i,j}) \end{array} \right\} \cdot \\ & \prod_{j \in \mathbf{J}} \sum_{x_{i,j}=0}^{S_{i,j}} P(X_{i,j} = x_{i,j}) \Bigg(\sum_{\mathbf{z}_{.,\mathbf{j-1}} \in \mathbf{\Omega}_{\mathbf{z}}^{\mathbf{j-1}}} P(\mathbf{Z}_{.,\mathbf{j-1}} = \mathbf{z}_{.,\mathbf{j-1}}) \cdot \\ & \Big(\sum_{\mathbf{r}_{.,.,\mathbf{j-1}} \in \mathbf{\Omega}_{\mathbf{r|z}}^{\mathbf{j-1}}} P(\mathbf{R}_{.,.,\mathbf{j-1}} = \mathbf{r}_{.,.,\mathbf{j-1}} | \mathbf{z}_{.,\mathbf{j-1}}) \cdot \\ & P(\mathbf{Y}_{.,\mathbf{j-1}} = \mathbf{z}_{.,\mathbf{j}} + \mathbf{y}_{.,\mathbf{j}} - \mathbf{x}_{.,\mathbf{j}} - \mathbf{r}'_{.,.,\mathbf{j-1}} \cdot \mathbf{I}_{I \times 1} \mid \mathbf{z}_{.,\mathbf{j-1}}) \Big) \Bigg). \end{aligned}$$

Proof. For $j = 1$, by Equation (3.2) and the independence of patient flows across stations in the first slot, we have

$$\begin{aligned} P(\mathbf{Y}_{.,1} = \mathbf{y}_{.,1} | \mathbf{Z}_{.,1} = \mathbf{z}_{.,1}) \;\; &= \;\; \prod_{i \in \mathbf{I}} P(Y_{i,1} = y_{i,1} | X_{i,1} = y_{i,1} + z_{i,1}) \\ &= \;\; \prod_{i \in \mathbf{I}} P(max(y_{i,1} + z_{i,1} - L_i, 0) = y_{i,1} | X_{i,1} = y_{i,1} + z_{i,1}) \\ &= \;\; \prod_{i \in \mathbf{I}} \left\{ \begin{array}{c} 1(y_{i,1} = 0) P(L_i \geqslant z_{i,j}) \\ + \\ 1(y_{i,1} > 0) P(L_i = z_{i,j}) \end{array} \right\}. \end{aligned}$$

For $j > 1$, using the Bayes rule, we can rearrange the conditional joint overflow

as

$$
\begin{aligned}
P(\mathbf{Y}_{.,j} &= \mathbf{y}_{.,j} | \mathbf{Z}_{.,j} = \mathbf{z}_{.,j}) = P(\mathbf{A}_{.,j} = \mathbf{y}_{.,j} + \mathbf{z}_{.,j} | \mathbf{Z}_{.,j} = \mathbf{z}_{.,j}) \\
&= \frac{P(\mathbf{Z}_{.,j} = \mathbf{z}_{.,j} | \mathbf{A}_{.,j} = \mathbf{a}_{.,j}) \cdot P(\mathbf{A}_{.,j} = \mathbf{a}_{.,j})}{\sum_{\mathbf{z}_{.,j} \in \Omega_z^j} P(\mathbf{Z}_{.,j} = \mathbf{z}_{.,j})} \\
&= \frac{P(\mathbf{A}_{.,j} = \mathbf{a}_{.,j})}{\sum_{\mathbf{z}_{.,j} \in \Omega_z^j} P(\mathbf{Z}_{.,j} = \mathbf{z}_{.,j})} \prod_{i \in \mathbf{I}} \left\{ \begin{array}{c} 1(y_{i,j} = 0) P(L_i \geqslant z_{i,j}) \\ + \\ 1(y_{i,j} > 0) P(L_i = z_{i,j}) \end{array} \right\}.
\end{aligned}
\tag{3.14}
$$

Substituting Equations (3.11) and (3.10) and Proposition 3.4.2 into Equation (3.14) yields Proposition 3.4.3 for $j > 1$. □

In summary, to evaluate the cost of a schedule, we need the joint overflow distribution $P(\mathbf{Y}_{.,\mathbf{j}})$ for all time slots. As shown in Proposition 4.3.3 and Figure 3.2, the joint overflow for each time slot depends recursively on the joint referral distribution and the conditional overflow distribution from an earlier slot. The conditional overflow depends on the joint distribution of the completed services. Essentially, the dimensionality of the joint distributions, together with the recursive structure, dictates the complexity of the model.

3.5. Sequential Scheduling Under A Coordinated Myopic Policy

In sequential scheduling, each patient contacts the network scheduler for an appointment and is offered a decision by the end of the call. Let $\mathbf{S^n}$ denote a schedule during the call-in process with n patient currently scheduled for arrival on the appointment day. Suppose that the next patient requests station i and is assigned to

slot j. The decision is denoted by $\Delta_{i,j}$, an $I \times J$ matrix with value 1 in cell (i, j) and 0 elsewhere. Let $\Delta_{i,0}$ be a zero matrix representing a rejection for the focal period. Upon each patient call, the scheduler's objective can then be denoted as

$$\max_{\Delta_{i,j},\, j \in \mathbf{J} \cup \{0\}} \{V_T(\mathbf{S^n} + \Delta_{i,j})\}. \tag{3.15}$$

The myopic optimal decision is denoted as j^* and the schedule evolves as $\mathbf{S^{n+1}} = \mathbf{S^n} + \Delta_{i,j^*}$. Figure 3.3 provides the scheduling mechanism. The centralized scheduler starts with an empty schedule book, an $I \times J$ zero matrix, $\mathbf{S^0}$. The scheduler also keeps track of the station-slot pairs that have been deemed as decreasing the objective, in the set $\mathbf{\Psi}$.

A call arrives requesting an appointment for a particular day, T. The scheduler obtains the patient's slot preferences, denoted by $\mathbf{C}$, ($\mathbf{C} = \mathbf{J}$ for availability anytime) and determines the no-show probability of the caller based on the requested service. The no-show probability can potentially be estimated from historical data. The algorithm finds the best slot $j^* \in \mathbf{C}$, provided $j^* \notin \mathbf{\Psi}$ and assigns the patient to it. If no such j^* exists, the patient is not offered a slot for that specific day and is considered for another. The schedule terminates when $\mathbf{\Psi}(i) = \mathbf{J}$, $\forall i \in \mathbf{I}$.

Step 1: Set $\boldsymbol{S^0} = \boldsymbol{0_{I\times J}}$, where $\boldsymbol{0_{I\times J}}$ is an $I \times J$ null matrix
$temp\boldsymbol{S^0} = \boldsymbol{0_{I\times J}}$
$\mathrm{V_T}(\boldsymbol{S^0}) = 0$
$n = 0$
$\boldsymbol{\Psi} = \emptyset$
Step 2: If $\boldsymbol{\Psi}(i) == \boldsymbol{J}, \forall\, i \in \boldsymbol{I}$, terminate the algorithm
Else, go to step 3
Step 3: Wait for the next patient call
Step 4: A patient calls and requests an appointment in period T at station i
The patient's no-show probability is ρ_i and his preferred slots are in set $\mathbf{C}$
If $\boldsymbol{C} \subseteq \boldsymbol{\Psi}$(i), reject the request for period T and go to step 3
Else, continue to step 5
Step 5: For each $j \in \boldsymbol{C}\backslash\{\boldsymbol{C} \cap \boldsymbol{\Psi}(\mathrm{i})\}$
Set $temp\boldsymbol{S}^{n+1}(j) = \boldsymbol{S}^n + \Delta_{i,j}$
Evaluate $\mathrm{V_T}(\mathrm{temp}\mathbf{S}^{n+1}(j))$
Step 6: If $\max[\mathrm{V_T}(\mathrm{temp}\mathbf{S}^{n+1}(j)), \forall\, j \in \boldsymbol{C}] > \mathrm{V_T}(\boldsymbol{S}^n)$
Set $j^* = argmax[\mathrm{V_T}(\mathrm{temp}\mathbf{S}^{n+1}(j)), \forall\, j \in \boldsymbol{C}]$
$Policy(n) = j^*$
Update $\boldsymbol{S}^{n+1} = \boldsymbol{S}^n + \Delta_{i,j^*}$
$n = n + 1$
Else, update $\boldsymbol{\Psi}(\mathrm{i}) = \boldsymbol{\Psi}(\mathrm{i}) \cup \mathbf{C}$
Go to step 2

Figure 3.3: Implementation lay out (pseudo-code)

3.5.1 A Note on Implementation

To implement this scheduling algorithm in real-time, a decision needs to be made during a patient call. This requires an efficient algorithm that can solve the scheduling problem in reasonable time. A straightforward approach is direct computation, that is, everything is computed when needed. Direct computation requires minimal memory storage. Moreover, due to the nonlinearity of the cost function, almost no previous calculations can be reused in the evaluation of a new schedule. This would result in unacceptably long computation times for patients waiting to

be scheduled over the phone. For example, for a 3-station, 8-slot network, direct computation can take over several hours. On the other extreme is complete pre-computation. We would pre-calculate and store the values of all possible schedules prior to patient calls. Pre-computation minimizes the time patients wait on the phone, but it consumes an intensive amount of preprocessing time and memory. For example, for an I-station, J-slot network with Q types of patient no-shows and u patients in a slot, the total number of schedules to be enumerated beforehand is $(u+1)^{I \cdot J \cdot Q}$, most of which are unlikely to be useful. For a 3-station, 8-slot network with two types of no-show probabilities, the number of schedules is more than 2.8 $\times 10^{14}$ and its pre-computation time is estimated to exceed 10^8 hours.

Leveraging on the advantages of each approach above, we propose a hybrid method that takes a significant yet necessary portion of computation offline. Due to patient no-shows, the actual number of patients showing up for their appointment is a random variable. A realization of a schedule is the actual number of patients that show up for their appointment, denoted as $\mathbf{O}^m$, $m \in M_S$. $M_{\mathbf{S}}$ is the set of all realizations of $\mathbf{S}$. For example, in a 2-station, 4-slot network, the schedule $\begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$ has four possible realizations on the appointment day: $\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$, $\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$, $\begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$ or $\begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$, with probabilities that depend on the patients' no-show probabilities. The value of a schedule is the expectation over all its realizations. Let $W(\mathbf{O}^m)$ denote the value of a realization $\mathbf{O}^m$. Then,

$$V_T(\mathbf{S}|\rho) = \sum_{m \in M_{\mathbf{S}}} P(\mathbf{O}^m|\mathbf{S},\rho) \cdot W(\mathbf{O}^m), \tag{3.16}$$

where $P(\mathbf{O}^m|\mathbf{S},\rho)$ is the probability that $\mathbf{O}^m$ is the realized schedule. In addition, we

calculate the value of a realization based on a 0% no-show rate. In this case, there is an one-to-one mapping from a schedule to its realization, and they are identical (i.e., $W(\mathbf{O}^m) = V_T(\mathbf{O}^m|\rho = 100\%)$).

In our implementation, we can precompute and store the values of all possible realizations, $W(\mathbf{O}^m)\,\forall m$, totaling $(u+1)^{I \cdot J}$ calculations using Equations (3.5) and (3.6). When scheduling, Step 5 of Figure 3.3 is simplified to Equation (3.16), which only involves look-up and integration. Instead of calculating the value of each schedule from scratch as in direct computation, our method utilizes the preprocessed $W(\mathbf{O}^m)$. Compared to the complete pre-computation, our method is efficient, as we only evaluate schedules that are visited in the construction of a schedule.

Our hybrid implementation method leverages on pre-computation to shift some computation time from during-the-call to before-the-call. However, as will be shown in Section 3.7, for even an 3-station, 8-slot network, the pre-computation part of the hybrid method requires over 15 hours on a powerful multi-core workstation. Such computational time may be prohibitively long for a healthcare practice with a common PC. To further reduce computation time, we design a series of approximation schemes in the next section. Both are necessary for successful implementation.

3.6. Approximation Schemes

The model includes three sources of uncertainties: the uncertainty that patients show up, the uncertainty of service time and patient routing between services. Each of these factors contributes to computational complexities. In this section, we ask

how each of these uncertainties affects the computational time and we seek an approximation scheme that gives significant computational advantages with good approximation quality. To this extent, we design three main approximation models, each replacing one of the three random variables by its expected value.

The first is to approximate patient no-show by its the expected attendance at each slot, rounded to the nearest integer, denoted as Model $\overline{X}$. In the second model, $\overline{Z}$, we replace stochastic service time by its mean. The third model substitutes the referral variable by its expected value, rounded to the nearest integer, denoted as $\overline{R}$. In our scheduling implementation, Model $\overline{X}$ only affects computation during the phone call, while the other two reduce pre-computation times. In this section, we only describe the three main approximations. In Section 3.7, we also combine these to form a series approximation models, i.e., Models $\overline{X}$, $\overline{Z}$, $\overline{R}$, $\overline{XZ}$, $\overline{XR}$, $\overline{ZR}$ and $\overline{XZR}$, and report the results. All approximations are applicable to any implementation of the scheduling algorithm.

3.6.1 Deterministic patient arrivals (Model $\overline{X}$)

Due to patient no-shows, we approximate the number of arrivals by its expectation, $\mathbf{X}_{.,j} \approx [\rho_1 \rho_2 ... \rho_I]'. \times \mathbf{S}_{.,j}$, where the operator $.\times$ is the element-wise multiplication of two vectors of the same size. Fractional numbers are rounded to the nearest integer. Thus, the joint overflow distribution is approximated by the following ex-

pression

$$
\begin{aligned}
P(\mathbf{Y}_{.,\mathbf{j}} = \mathbf{y}_{.,\mathbf{j}}) \;\approx\; & \sum_{\mathbf{a}_{.,\mathbf{j}} \in \Omega_a^j} \prod_{i \in \mathbf{I}} \left\{ \begin{array}{c} 1(y_{i,j} = 0) \cdot P(L_i \geq a_{i,j} - y_{i,j}) \\ + \\ 1(y_{i,j} > 0) \cdot P(L_i = a_{i,j} - y_{i,j}) \end{array} \right\} \cdot \\
& \Bigg(\sum_{\mathbf{z}_{.,\mathbf{j-1}} \in \mathbf{\Omega_z^{j-1}}} P(\mathbf{Z}_{.,\mathbf{j-1}} = \mathbf{z}_{.,\mathbf{j-1}}) \cdot \\
& \Bigg(\sum_{\mathbf{r}_{.,.,\mathbf{j-1}} \in \mathbf{\Omega_{r|z}^{j-1}}} P(\mathbf{R}_{.,.,\mathbf{j-1}} = \mathbf{r}_{.,.,\mathbf{j-1}} |\, \mathbf{Z}_{.,\mathbf{j-1}} = \mathbf{z}_{.,\mathbf{j-1}}) \cdot \\
& P(\mathbf{Y}_{.,\mathbf{j-1}} = \mathbf{a}_{.,\mathbf{j}} - \rho \times \mathbf{S}_{.,\mathbf{j}} - \mathbf{r}'_{.,.,\mathbf{j-1}} \cdot \mathbf{I}_{I \times 1} \mid \mathbf{Z}_{.,\mathbf{j-1}}) \Bigg) \Bigg).
\end{aligned}
$$

Note here that the arrival distribution is gone and the arrival variable is replaced by a constant in the conditional overflow distribution. Equation (3.9) can be written as $U(i,j) = \sum_{i \in \mathbf{I}} \sum_{j' < j} \rho_i S_{i,j'} + \rho_i S_{i,j}$, and the corresponding support set shrinks.

3.6.2 Deterministic service time (Model $\overline{Z}$)

Here, we remove the uncertainty in service time. As we approximate the service time by its mean, the number of completed services given a non-empty queue becomes $L_i \approx \lambda_i$, which refines the support set for $\mathbf{A}_{.,\mathbf{j}}$. Recall from Equation (3.3), the number of completed services is now approximated by $Z_{i,j} \approx \min(\lambda_i,\, A_{i,j})$. As a

result, the joint overflow distribution becomes

$$
\begin{aligned}
P(\mathbf{Y}_{.,\mathbf{j}} = \mathbf{y}_{.,\mathbf{j}}) \;\approx\; & \sum_{\mathbf{a}_{.,\mathbf{j}} \in \Omega_a^j} \prod_{i \in \mathbf{I}} \left\{ \begin{array}{c} 1(y_{i,j} = 0) \cdot 1(a_{i,j} \leq \lambda_i) \\ + \\ 1(y_{i,j} > 0) \cdot 1(a_{i,j} = y_{i,j} + \lambda_i) \end{array} \right\} \cdot \\
& \prod_{j \in \mathbf{J}} \sum_{x_{i,j}=0}^{S_{i,j}} P(X_{i,j} = x_{i,j}) \Bigg(\sum_{\mathbf{z}_{.,\mathbf{j-1}} \in \mathbf{\Omega_z^{j-1}}} P(\mathbf{Z}_{.,\mathbf{j-1}} = \mathbf{z}_{.,\mathbf{j-1}}) \cdot \\
& \Bigg(\sum_{\mathbf{r}_{.,.,\mathbf{j-1}} \in \mathbf{\Omega_{r|z}^{j-1}}} P(\mathbf{R}_{.,.,\mathbf{j-1}} = \mathbf{r}_{.,.,\mathbf{j-1}} |\, \mathbf{Z}_{.,\mathbf{j-1}} = \mathbf{z}_{.,\mathbf{j-1}}) \cdot \\
& P(\mathbf{Y}_{.,\mathbf{j-1}} = \mathbf{a}_{.,\mathbf{j}} - \mathbf{x}_{.,\mathbf{j}} - \mathbf{r}'_{.,.,\mathbf{j-1}} \cdot \mathbf{I}_{I \times 1} \mid \mathbf{Z}_{.,\mathbf{j-1}}) \Bigg) \Bigg).
\end{aligned}
$$

In comparison to the expression in Proposition 3.4.1, the indicator functions can be viewed as additional constraints on Ω_a^j. Moreover, the joint service completion distribution in Proposition 3.4.2 is approximated as

$$
\begin{aligned}
P(\mathbf{Z}_{.,\mathbf{j}} = \mathbf{z}_{.,\mathbf{j}}) \;\approx\; & \sum_{\mathbf{a}_{.,\mathbf{j}} \in \mathbf{\Omega_a^j}} \prod_{i \in \mathbf{I}} \left\{ \begin{array}{c} 1(z_{i,j} < \lambda_i) \cdot 1(a_{i,j} < \lambda_i) \\ + \\ 1(z_{i,j} = \lambda_i) \cdot 1(a_{i,j} \geq \lambda_i) \end{array} \right\} \cdot \\
& \prod_{i \in \mathbf{I}} \sum_{x_{i,j}=0}^{S_{i,j}} P(X_{i,j} = x_{i,j}) \Bigg(\sum_{\mathbf{z}_{.,\mathbf{j}-1} \in \mathbf{\Omega_z^{j-1}}} P(\mathbf{Z}_{.,\mathbf{j-1}} = \mathbf{z}_{.,\mathbf{j-1}}) \cdot \\
& \Bigg(\sum_{\mathbf{r}_{.,.,\mathbf{j}-1} \in \mathbf{\Omega_{r|z}^{j-1}}} P(\mathbf{R}_{.,.,\mathbf{j-1}} = \mathbf{r}_{.,.,\mathbf{j-1}} |\, \mathbf{Z}_{.,\mathbf{j-1}} = \mathbf{z}_{.,\mathbf{j-1}}) \cdot \\
& P(\mathbf{Y}_{.,\mathbf{j-1}} = \mathbf{a}_{.,\mathbf{j}} - \mathbf{x}_{.,\mathbf{j}} - \mathbf{r}'_{.,.,\mathbf{j-1}} \cdot \mathbf{I}_{I \times 1} \mid \mathbf{Z}_{.,\mathbf{j-1}} = \mathbf{z}_{.,\mathbf{j-1}}) \Bigg) \Bigg).
\end{aligned}
$$

Again, the indicator functions shrink the support set for Ω_a^j. A drawback here is that it over-estimates the number of completed services. Only a finite number of patients is in any queue, so the completed service follows a truncated Poisson distribution,

and its expectation is smaller than the service rate. Therefore, Model $\overline{Z}$ tends to accommodate more patients in the schedule.

3.6.3 Deterministic referral routing (Model $\overline{R}$)

Here, we assume that the number of referrals is proportional to the completed services: $R_{i,k,j} \approx P_{i,k} \cdot Z_{i,j}\ \forall i,k \in \mathbf{I}\ j \in \mathbf{J}$. Fractional numbers are rounded to the nearest integer. So the number of referrals received at slot $i+1$ can be expressed as: $\mathbf{P}' \cdot \mathbf{Z}_{.,\mathbf{j}}$. As a result, the joint overflow distribution is approximated as

$$
\begin{aligned}
P(\mathbf{Y}_{.,\mathbf{j}} = \mathbf{y}_{.,\mathbf{j}}) \approx & \sum_{\mathbf{a}_{.,\mathbf{j}} \in \Omega_a^j} \prod_{i \in \mathbf{I}} \left\{ \begin{array}{c} 1(y_{i,j}=0) \cdot P(L_i \geq a_{i,j} - y_{i,j}) \\ + \\ 1(y_{i,j}>0) \cdot P(L_i = a_{i,j} - y_{i,j}) \end{array} \right\} \cdot \\
& \prod_{j \in \mathbf{J}} \sum_{x_{i,j}=0}^{S_{i,j}} P(X_{i,j} = x_{i,j}) \Big(\sum_{\mathbf{z}_{.,\mathbf{j-1}} \in \mathbf{\Omega_z^{j-1}}} P(\mathbf{Z}_{.,\mathbf{j-1}} = \mathbf{z}_{.,\mathbf{j-1}}) \cdot \\
& \cdot P(\mathbf{Y}_{.,\mathbf{j-1}} = \mathbf{a}_{.,\mathbf{j}} - \mathbf{x}_{.,\mathbf{j}} - \mathbf{P}' \cdot \mathbf{z}_{.,\mathbf{j-1}} \mid \mathbf{Z}_{.,\mathbf{j-1}} = \mathbf{z}_{.,\mathbf{j-1}}) \Big).
\end{aligned}
$$

This approximation saves the integration of $\mathbf{r}_{.,.,\mathbf{j-1}}$, which is the inner loop in the $P(\mathbf{Y}_{.,\mathbf{j}} = \mathbf{y}_{.,\mathbf{j}})$ expression. So the time saved using this method is expected to be less than that of Model $\overline{Z}$.

3.7. Computational Examples and Insights

Our multi-station model and scheduling mechanism are developed for a general network and can be applied in many different healthcare and general service settings. We choose an example network that is simple enough to generate insights and complex enough to capture patient routing and the scheduling aspects of the

network. As shown in Figure 3.4, the network consists of three stations that are connected via the stochastic referral routing shown by the arrows. Patients enter the network at station 1 or 3 and station 2 only accepts internal referrals from station 1. A fraction of station 1's patients require additional services at stations 2 and 3. All patients who complete service at station 2 proceed to station 3 for further examination. This network structure is very common in primary care clinics, where Station 1 represents a general physician, Station 2 signifies a lab or X-ray, and Station 3 denotes a specialist. Such structures are also common in inter-clinic referrals. For example, in pre-operative outpatient surgical care, an anesthesiologist (Station 1) may refer patients to labs (Station 2) and to other specialists such as internal medicine (Station 3) in order to optimize a patient for surgery.

In this section, we use this network because of its elegance, applicability and ability to capture all the necessary elements and managerial insights we would like to convey. For the sake of clearer insights and chapter length, we have refrained from an exhaustive set of computational examples on more complex networks, although every element of the chapter discussed until now are directly applicable. In what follows, we demonstrate the construction of a network schedule, evaluate the performance of the coordinated myopic policy relative to an unachievable super-optimal policy, seek insights into the dependence on the model parameters, and finally examine the performance of the approximation schemes proposed in Section 3.6.

We first fix a base case parameter choice and then consider modifications to this base set when needed. Suppose that 50% of station 1's patients leave the net-

work after service; while $P_{1,2} = 25\%$ and $P_{1,3} = 25\%$ of station 1's patients require additional services at stations 2 and 3, respectively. Every station operates 8 slots with the same service rate, which is normalized to 1. For each service provided, we assume the network receives a nominal reward of 100, regardless of the station, i.e., $r_1 = r_2 = r_3 = 100$, and all costs are defined relative to the reward. Due to probabilistic routing, the expected reward collected from a patient depends on the initial service requested, which can be calculated using Equation (3.5): $R_1 = r_1 + P_{1,2}(r_2 + P_{2,3}r_3) + P_{1,3}r_3$, $R_2 = r_2 + P_{2,3}r_3$ and $R_3 = r_3$. We monetize the penalty for patient waiting and station overtime as $\mathbf{c_j} = 0.25\mathbf{r}$, $\forall j \neq J$ and $\mathbf{c_J} = 1.5\mathbf{R}$. As $\mathbf{c_j}\,(j \neq J)$ measures the relative cost of an overflow to the reward of a completed service, it should be less than 1. On the other hand, $\mathbf{c_J}$ should be greater than 1 because serving a patient during clinic overtime costs more than the reward. Other coefficients besides the 0.25 and 1.5 for the $\mathbf{c_j}\,(j \neq J)$ and $\mathbf{c_J}$, will be considered in a sensitivity analysis in Section 3.7.3. Each time a patient calls, there is equal probability that the request is for station 1 or 3. All patients scheduled have a 60% chance of attending their appointments on that day, i.e., $\rho_1 = \rho_3 = 60\%$.

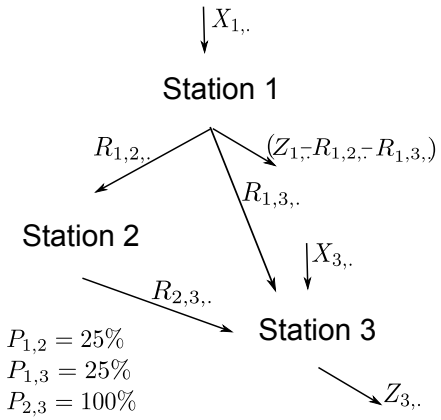

Figure 3.4: Example Network

All computations were performed on a linux Dell Poweredge 2950 workstation with 2 six-core, hyperthreading 3.33 GHz Xeon processors and 24 GB of shared memory. The time reported is the wall-time of the machine.

3.7.1 Construction of a Schedule

We present here an illustration of a network schedule being constructed sequentially. We use a randomly generated sequence of requests shown in Table 3.1. Columns 1 and 2 represent the patient number and the station to which the respective patient asked to be scheduled. The slot for which our policy schedules them is shown in column 3. When the schedule is relatively packed, we can see that the policy starts reject all patients requesting station 3 while still accepting patients for station 1. Once station 1 becomes packed as well, the scheduling stops, with patient 22. The network objective value is shown in column 4 and is broken down into the expected revenue and costs in the last two columns. The final schedule is shown

Slot	1	2	3	4	5	6	7	8
Station 1	②⑫	⑤⑳	④	⑩	⑧	⑭	⑰	
Station 3	①⑪	⑥	⑦	⑮	③	⑨	⑬	

Figure 3.5: An Example of the Decision Process

in Figure 3.5, with nine patients scheduled for station 1 and eight for station 3. The numbers indicate the patient number. For example, patient numbers 2 and 12 are the ones that requested for station 1 and were scheduled to slot 1 of station 1. Since station 2 does not accept exogenous patient requests, it is omitted in the expression of a schedule.

Patient Sequence	Request Type	Slot Decision	Expected Profit	Expected Revenue	Expected Cost
1	3	1	51.25	60	8.75
2	1	1	139.45	165	25.55
3	3	5	187.79	225	37.21
4	1	3	267.99	330	62.01
5	1	2	340.76	435	94.24
6	3	2	382.79	495	112.21
7	3	3	416.59	555	138.41
8	1	5	479.03	660	180.97
9	3	6	504.54	720	215.46
10	1	4	551.68	825	273.32
11	3	1	573.23	885	311.77
12	1	1	610.95	990	379.05
13	3	7	619.41	1050	430.59
14	1	6	646.55	1155	508.45
15	3	4	648.39	1215	566.61
16	3	-	648.39	1215	566.61
17	1	7	658.26	1320	661.74
18	3	-	658.26	1320	661.74
19	3	-	658.26	1320	661.74
20	1	2	662.70	1425	762.30
21	3	-	662.70	1425	762.30
22	1	-	662.70	1425	762.30

Table 3.1: Evolution of Network Schedule

Notice that the scheduling for any station begins by distributing patients across that day and then moves toward packing them in more tightly. For example, station 3's first and second requests are scheduled to slots 1 and 5, and the next request comes back to slot 2. With a 40% no-show rate, station 1 double-books in the first two slots, while station 3 only double-books its first slot, in anticipation of referrals from station 1. Also notice that both stations leave the last slot open to cope with overflows in later slots so as to avoid the overtime penalty. This practice is very common in clinics because schedulers have intuitively learned to accommodate the

significant cost of overflow at the end of the day.

3.7.2 Optimality Gap

In this subsection, we address the question of how good the myopic policy is. Previously, we have pointed out that a non-myopic policy would suffer from insurmountable complexity, along with extreme sensitivity to parameters. Although myopic polices do not have these issues, the question of how much we give up in terms of the objective value still must be raised. In fact, the context of clinical scheduling aligns well with the nature of a myopic policy. That is, scheduling decisions are critical only when a schedule is well packed but very forgiving when the schedule is sparse. Myopic polices indeed yield good approximation towards the end of the scheduling horizon, when schedules are packed.

Quantifying the quality of a myopic policy is a challenging task, especially because we cannot contrast it to the optimal policy, which is non-computable for the reasons discussed earlier. However, we can compute a super-optimal policy which is unachievable. The difference between the myopic and this super-optimal value provides an upper bound for the actual optimality gap.

The super-optimal schedule is the schedule that maximizes profits by assigning patients into any possible slots at the morning of the appointment, assuming unlimited requests of any type. The computation of this policy, like the myopic, does not depend on models for call-in arrivals, and is therefore unachievable for certain call-in sequences. Let the super-optimal value be denoted as V_{so}. The super-optimality

gap (SOG) is defined as

$$SOG = \sum_{n} \frac{V_{so} - V^{n}_{coord.}}{V_{so}}. \tag{3.17}$$

Here and in subsection 3.7.3, we run the coordinated myopic policy on 2000 randomly generated call-in sequences, which is found to be statistically sufficient for our numerical analysis. At the end of each simulation run, we obtain the value of the schedule, defined as $V^{n}_{coord.}$, and the number of patients scheduled.

Finding the super-optimal solution requires an exhaustive search through the entire state space of all possible schedules. Due to the challenge of high dimensionality, we resort to a 4-slot network in computing the SOGs for a range of parameter settings: cost of overflow, cost of overtime, referral probabilities and no-show rates. The design of experiments and the corresponding SOGs are shown in Table 3.2.

c_j	$SOG\,(\%)$
$2/8r$	0.18
$3/8r$	0.04
$4/8r$	0.57
$5/8r$	0.00
$6/8r$	0.98

$\mathbf{c_J}$	$SOG\,(\%)$
$10/8\mathbf{R}$	0.11
$11/8\mathbf{R}$	0.38
$12/8\mathbf{R}$	0.18
$13/8\mathbf{R}$	0.51
$14/8\mathbf{R}$	0.53

$P_{1,2}, P_{1,3}$	$SOG\,(\%)$
$0.25, 0.25$	0.18
$0.25, 0.50$	0.58
$0.25, 0.75$	0.91
$0.50, 0.25$	0.43
$0.50, 0.50$	0.70
$0.75, 0.25$	0.28

ρ_1, ρ_2	$SOG\,(\%)$
$0.6, 0.6$	0.18
$0.6, 0.8$	0.24
$0.6, 1.0$	0.37
$0.8, 0.6$	0.25
$0.8, 0.8$	0.12
$0.8, 1.0$	0.29
$1.0, 0.6$	0.70
$1.0, 0.8$	0.87
$1.0, 1.0$	0.53

Table 3.2: Super Optimality Gap

As Table 3.2 reveals, the coordinated myopic solutions are within 1.0% of the super-optimal bound on all the parameter values we test. As the actual optimal

value is lower than the super-optimal bound, we expect the actual optimality gap to be smaller.

3.7.3 Dependence on Model Parameters

In this subsection, we ask the question of how the network objective value changes with respect to various model parameters and investigate the underlying reasons.

We discuss these directional insights observed from the 8-slot model using the coordinated myopic policy. Figure 3.6 is a visual summary of the changes in $\overline{V}(S)$ as the value of the parameter varies. As expected, when the cost coefficients increase, $\overline{V}(S)$ decreases. Relatively speaking, the coefficient for $\mathbf{c_J}$ has a smaller effect on the solution value because it only affects the last slot.

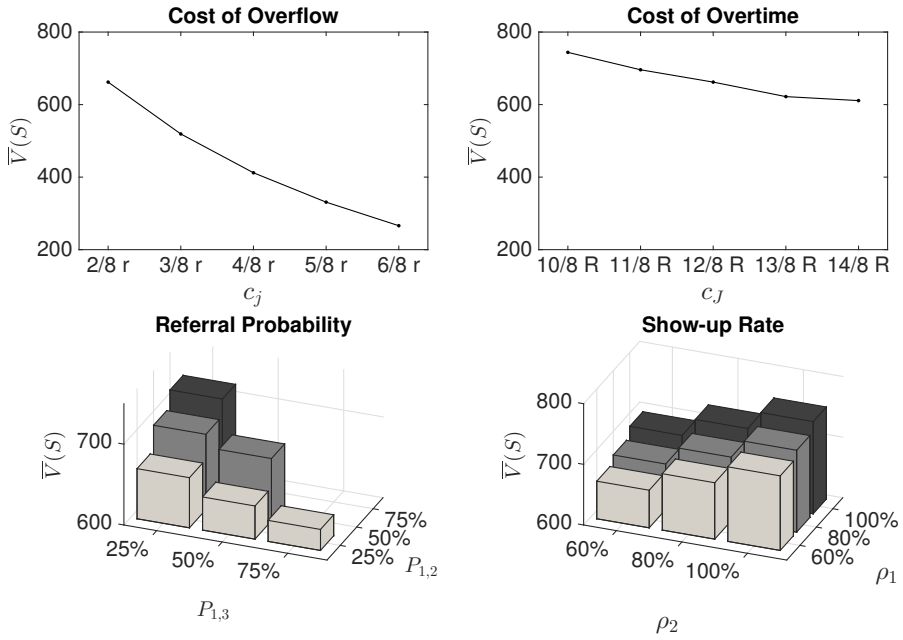

Figure 3.6: Sensitivity Analysis

Next, we hold $P_{2,3} = 100\%$ and examine the interplay between $P_{1,2}$ and $P_{1,3}$. At any level of $P_{1,3}$, a higher rate of $P_{1,2}$ boosts $\overline{V}(S)$; while at any level of $P_{1,2}$, the trend reverses for increasing $P_{1,3}$. The reason lies in the interaction between utilization of the stations and the profit margin of different requests, as $P_{1,2}$ or $P_{1,3}$ changes. Recall that patients who visit station 2 proceed to station 3. Therefore, a higher $P_{1,2}$ implies higher expected revenue per type 1 patient and a higher utilization of station 2. Since station 2 is the least busy resource, its cost for an additional patient is the smallest among all stations. Therefore, the profit margin per type 1 patient increases in $P_{1,2}$, as does the value of the schedule. On the other hand, higher $P_{1,3}$ significantly increases the utilization of station 3, for which the cost of congestion

outweighs the moderate increase in the revenue of type 3 patients. Therefore, $\overline{V}(S)$ decreases in $P_{1,3}$. The key insight is that when a referral rate changes in the direction toward a more balanced network in terms of resource utilization, the objective value increases. Lastly, our results on no-show align with the literature: The lower the uncertainty in no-show, the better a clinic's performance is.

3.7.4 Performance of Approximation Schemes

For the approximation schemes described in Section 3.6, we next compare them to the coordinated myopic solution on a range of parameter settings. We discuss computational efficiencies based on our hybrid algorithm implementation described in Section 3.5.1. Specifically, we present the pre-computation times and the online computation times for each approximation. The purpose of this section is also to shed light on important factors practitioners need to consider to coordinate appointment scheduling in their own healthcare network.

We measure the performance of each approximation method by its relative difference to the coordinated method: $\frac{\overline{V}_{coord.}-\overline{V}_{appr.}}{\overline{V}_{coord.}}$. Again, we run the approximation methods on the same 2000 call-in sequences, and the value is denoted as $V^n_{appr.}$, where the subscript refers to the corresponding approximation method. The estimated value function by any of the heuristics is measured as $\overline{V}_{heuristic} = \frac{1}{2000}\sum_{n=1}^{2000} V^n_{heuristic}$, with n indicating the simulation run.

We start by contrasting the pre-computational times for various approximations against the non-approximated Coord. method, using the base case parameters. In Figure 3.7, the performance of each method is measured as the percentage differ-

ence to the Coord. solution value, presented on the Y-axis. The smaller the percentage difference on the Y-axis, the better the solution quality. As can be seen, pre-computation can be extremely expensive even for a 8-slot-3-station network. Compared to the Coord. method, Model $\overline{Z}$ alleviates the preprocessing workload by 92% (reducing offline time to 1.3 hours); while Model $\overline{R}$ reduces the workload by 78% (3.3 hours of offline computation). The combination of the two, $\overline{ZR}$, further reduces the computation time by a small amount. By construction, Model $\overline{X}$ requires the same amount of computation as the Coord. method and its variants (i.e., Model $\overline{XZ}$, $\overline{XR}$ and $\overline{XZR}$) require the same amount of computation as the non $\overline{X}$ part. In terms of solution values, methods that approximates no-show all yield poor solution.

The Coord. method demands over 15 hours of preprocessing on a 32-core workstation running in parallel. However, when relying on a more common desktop, the actual offline time will be significantly longer. Furthermore, for larger clinical networks, the Coord. method may not even be feasible at all, even on high-performance workstations. These issues necessitate the development of the approximation methods.

Next, we turn to the computations involved during the scheduling process. The hope is to be able to accommodate these computational times when the patient is waiting on the call or at least between calls. The X-axis on Figure 3.8 marks the sequence of appointment decisions and their corresponding values on the Y-axis are the time it takes to make that decision. Because of rejections, the number of patients in the schedule may be smaller than the sequence number. A higher sequence

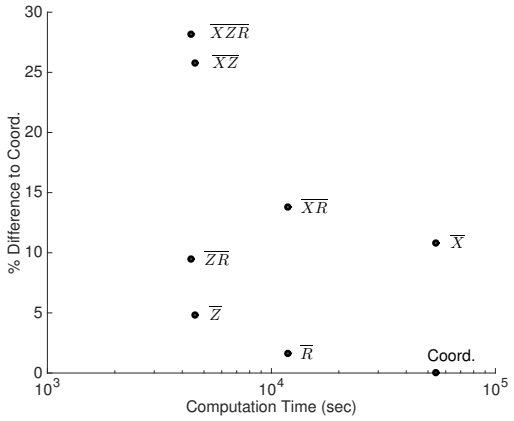

Figure 3.7: Offline Computation Time v.s. Solution Quality.

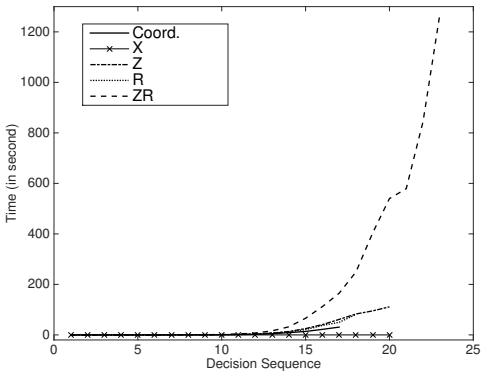

Figure 3.8: Online Computation Time per Decision.

number implies a more congested system. Based on Step 5 in Figure 3.3 and our implementation method described in Section 3.6, the determinant of time per decision is the computation complexity of Equation (3.16). Specifically, since the values of all realizations have been preprocessed, the computation time for Equation (3.16) depends only on the number of realizations associated with the schedule to be evaluated.

Because Model $\overline{X}$ approximates the value of a schedule using the expected number of arrivals, it simplifies Equation (3.16) to $V_T(\mathbf{S}) \approx W(E(\mathbf{S}|\rho))$. Therefore, the online time of Model $\overline{X}$ is independent of how many patients are already scheduled, shown as a flat line in Figure 3.8. In addition, due to rounding error, Model $\overline{X}$ tends to schedule excessively at a high no-show rate but to under-book when the no-show rate is low. With our no-show rate, we observe a more congested schedule under

Model $\overline{X}$ than the Coord. method. For the same reason as Model $\overline{X}$, all its variants exhibit a similar pattern, that is, close-to-zero decision time and congested schedule. Hence, we do not include them in Figure 3.8.

Except for Model $\overline{X}$ and its variants, all methods exhibit an exponential growth in online time as more patients join the network. The reason is that as the schedule fills up, the number of realizations to be integrated in the evaluation of each schedule (Equation (3.16)) increases exponentially. Moreover, depending on the resulting schedule under each policy, the total combinations of the possible realizations varies, and thus the evaluation time per decision differs among Coord., Model $\overline{Z}$, Model $\overline{R}$ and Model $\overline{ZR}$. For example, an evenly spread schedule like $\begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$ involves 2^3 realizations; while a schedule with the same number of patients but with a double-booking in the first slot, i.e., $\begin{bmatrix} 2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$, involves only 6 realizations. So the earlier a policy starts over-booking, the less online time it takes to offer a decision. The impact of deterministic service time, as in Model $\overline{Z}$, is to evenly distribute appointment requests among slots, before any over-booking. In contrast, the Coord. method favors certain slots over others and it leaves gaps between occupied slots and double-books early. The impact of deterministic patient routing is in-between Coord. and Model $\overline{Z}$. Their combined method, Model $\overline{ZR}$, spreads the appointment requests more evenly. Therefore, we observe that Model $\overline{ZR}$ takes the longest time to make a decision during later requests, followed by Model $\overline{Z}$, and Model $\overline{R}$.

In terms of the stopping point, Model $\overline{Z}$ books more patients than the Coord. solution because it over-estimates system capacity (discussed in Section 3.6.2). Model $\overline{R}$ may either over or under book relative to the Coord. method, which is de-

pendent on its referral rate. Due to rounding errors, infrequent referrals results an under-estimation of the network's congestion level and Model $\overline{R}$ books more than Coord.; the opposite occurs when referrals are more frequent. Model $\overline{ZR}$ amplifies the approximation errors from Model $\overline{Z}$ and Model $\overline{R}$ and significantly underestimates the network congestion. By the above reasoning, we observe the patterns in Figure 3.8, where Model $\overline{ZR}$ schedules excessively more patients.

Comparing among the approximation methods, Model $\overline{X}$ and its variants are the least appealing methods due to large approximation error and inefficient pre-processing. Despite its shortest pre-computation time, Model $\overline{ZR}$ is not practical online. The choice between Model $\overline{Z}$ and Model $\overline{R}$ should be made on a case-by-case basis. We recommend that healthcare management consider three main factors: (1) the desired solution quality, (2) computation resources for pre-computation and (3) patient tolerance for waiting on the phone. In a situation where network parameters are subject to frequent changes, Model $\overline{Z}$ is most favorable, with fast pre-computation and good solution quality on a wide range of parameters. The downside is longer waiting time for the later patients to get a decision. On the other hand, for a stable network with high demand for solution quality, Model $\overline{R}$ becomes attractive, especially when the cost of patient overflow is high and referrals infrequent.

Moving on to the dependence of these computational times on problem parameters, we summarize the results in Figure 3.9. The insights from Figure 3.9 are the following. First, as $\mathbf{c_j}$ increases, the quality of all approximations deteriorates but Model $\overline{R}$ yields the best result, followed by Model $\overline{Z}$. A similar pattern is observed

with $\mathbf{c_J}$, except for $\mathbf{c_J} = 1.25\mathbf{R}$, where Model $\overline{R}$ and Model $\overline{Z}$ are statistically indifferent. At different referral probabilities, the most robust method is Model $\overline{Z}$ with a solution 5%-7% lower than the Coord. method. In contrast to the robustness of Model $\overline{Z}$, solutions by Model $\overline{R}$ degrade as referral rate increases. This is expected because the impact of rounding errors is more severe when referrals are more frequent. In fact, at $P_{1,2} = P_{1,3} = 50\%$, Model $\overline{R}$ significantly overestimates the number of referrals out of clinic 1 and thus schedules too few people to fully utilize the network. The solution value is therefore significantly lower. Lastly, for different no-show probabilities, Model $\overline{Z}$ and Model $\overline{R}$ are both robust, and Model $\overline{R}$ is better. Model $\overline{X}$ only comes close to the performance of Model $\overline{R}$ when attendance rates are high (80%,80%).

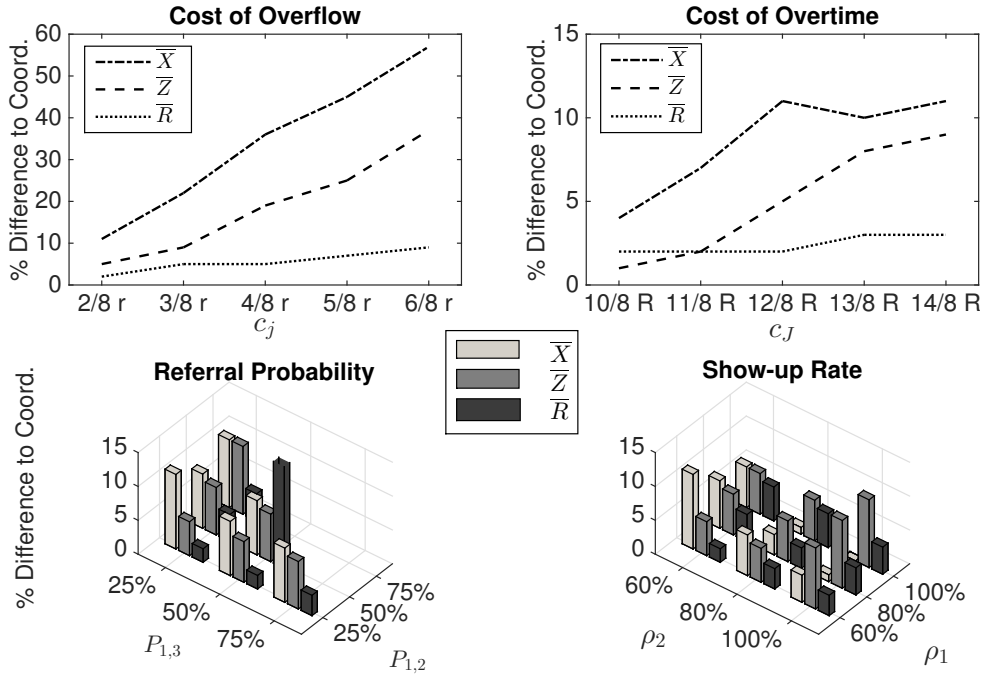

Figure 3.9: Quality Performance.

To summarize, even the myopic approach that has been proposed is computationally too complex for implementation. The computational times are too far from acceptable, especially when a patient is on a phone-call waiting for the result. Our hope was to find an approximation scheme that, together with the hybrid computational implementation described in Section 3.6 would allow a reasonable approximation with acceptable computational times. Our finding is that, among the approximations considered, we have a number that fall well within the acceptable threshold.

3.8. Concluding Remarks

The focus of current healthcare reforms revolves around improving efficiency and effectiveness of care. The entire industry is beginning to retool itself, all the way from redesigning hospital layouts to fundamentally changing billing structures, to improve coordination and move toward a patient-centered clinical model. With most patient care in the United States being delivered through outpatient clinics, coordinating appointment scheduling is arguably a key component of successful reform. This chapter provides the necessary technological contributions. The focus has been to construct a model formulation and to provide the analysis, insights and the computational tools necessary for real-world adoption. The model has carefully avoided assumptions that might have theoretical benefits but ultimately make the model unsuitable for implementation.

Note that our view of coordination is also limited to appointment scheduling in a multi-speciality hospital or multiple clinics in a common location. The emerging practice of coordinated, patient-centric care is broader. Examples include patient-centered medical homes in primary care (Stange et al. (2010)) and perioperative (patient-centered) surgical homes in outpatient surgery (Vetter et al. (2013, 2014); Morrice et al. (2014)). Such coordinated care is being recognized at world-class healthcare institutions, such as Virginia Mason Medical Center in Seattle, WA, and MD Anderson Cancer Center in Houston, TX. The newly established Dell Medical School at the University of Texas, Austin considers value-based care central to its mission. Although not straightforward, we believe that our work can be extended or adapted to these and other patient-centered models.

Chapter 4

Coordinated Scheduling for a Multi-server Network in Outpatient Surgical Care

4.1. Introduction

System integration and coordination are increasingly important in healthcare delivery, particularly in the United States with the new healthcare mandate. These concepts are vital for improving patient outcomes and satisfaction, optimizing providers' utilization and reducing operational costs in outpatient surgical care. As pointed out by Porter (2009), the overarching strategy of our healthcare reform should be focused on maximizing patient outcomes at the lowest cost, i.e. a value-based care, which is defined as health outcome over cost spending. To manage cost by improving coordination and operational efficiency is particularly critical in the surgical episode of care, because it involves multiple providers (e.g., primary care physicians, surgeons, anesthesiologists, and nurses) depending on patients' needs. Historically, these services have been fragmented with minimal coordination among providers. Consequentially, patients have to plan for and coordinate their own medical trips across multiple services. As a result of fragmented care, patients often experience long access delays to and between different services. For the same reason, providers frequently experience high fluctuations in their daily operations. Since outpatient surgeries account for the majority of all surgery visits in the United

States (Cullen et al. (2009), Berg and Denton (2012)), the potential impact for systems improvements and operational cost savings in this area is great.

The current healthcare reform aims to transform the industry from physician-centered to patient-centric. Reimbursement schemes are transitioning from volume-driven to value-based to incentivize coordination among different services to jointly deliver care in an effective and efficient manner (Health Cost Containment and Efficiencies (2010)). As a result, many patient-centric models are emerging, such as the Patient-Centered Medical Home (PCMH) (Stange et al. (2010)), the Perioperative Surgical Home (PSH) (ASA (2011)), and the Patient-Centered Surgical Home (PCSH) (Morrice et al. (2014)). In a PSH, care is integrated and coordinated among specialties, with anesthesiologists serving as system coordinators and information integrators in collaboration with the surgeons, general internists and other physicians (Vetter et al. (2013, 2014)). The PCSH is developed on the concept of PSH, with an emphasis on the centrality of improved patient care similar to the PCMH found in the practice of primary care. In particular, the PCSH model ensures close coordination between anesthesiologists and general internists in preoperative care so as to optimize patients' health conditions before their surgery. This approach has conceptual appeal because patient health problems discovered by an anesthesiologist during a surgery pre-assessment can be addressed by a general internist. Moreover, the PCSH aims to enable same-day referrals so that patients can see both providers, if needed, on a single visit.

A challenge facing the PCSH is how to coordinate the anesthesiologists and general internists who often reside in separate clinics. More specifically, it is common

for anesthesiologists to reside in an anesthesiologist preoperative clinic (APC) and general internists to be housed in an separate Internal Medicine Clinic (IMC). With this arrangement, it is typical for APC and IMC to schedule their own patients independently with no coordination. We refer to this type of scheduling policy as the Silo policy. Under Silo, same-day referrals can cause serious system congestions in the PCSH, especially at IMC. Since Silo can book IMC to its full capacity with no consideration for the referral patients, on the appointment day, providers at IMC either have to work the APC referrals into their already busy schedule, or defer them to another day. The former approach risks incurring a significant cost of patient waiting and clinical overtime, while the latter approach risks having patients abandon their IMC appointments and show up on the day of surgery with health issues that can lead to surgery delays or cancellations.

As an alternative to the Silo policy, IMC could leave empty blocks in its schedule to accommodate the expected increase in demand from APC referral patients on the appointment day. While this represents a level of coordination, it is static in nature because the empty blocks are predetermined. Hence, we refer to this policy as the Static policy. Because the blocked slots are predetermined, the providers at IMC still have to work the patients into the schedule, if referrals arrive at non-designated slots. As a result, although the Static policy has the potential to reduce system inefficiencies found in Silo, its effectiveness to coordinate APC and IMC may be limited.

To address potential problems associated with the Silo and the Static policies, we propose a fully coordinated scheduling approach that sequentially allocates ap-

pointment requests to both APC and IMC in order to optimize a PCSH "system" objective. This is a non-trivial exercise because APC and IMC operate with different number of providers that work in parallel. Furthermore, both services face uncertainties in patient attendance, service times and referrals.

We formulate a 2-clinic network model with multiple service providers that work in parallel in each clinic. The model accounts for the uncertainty of patient no-shows, stochastic service times and the inter and intra-clinic patient flows. The objective is to maximize the overall profit of the PCSH, which is the net of service rewards and the costs of patient waiting and clinical overtime. We propose a coordinated myopic policy that ensures a balanced utilization of clinical resources in the PCSH with high-quality, robust solutions. We develop a simulation-aided scheduling method to sequentially construct a schedule, using the statistical method of Ranking and Selection (R&S) (Law and Kelton (2010), Kim and Nelson (2001)). By using the R&S procedure, we are able to guarantee with confidence that the final schedule is sufficiently close to the best solution. Moreover, to improve the computation efficiency and the solution accuracy, we embed a hybrid evaluation method, which computes the profit of the schedule using a combination of analytical and simulation methods, in the scheduling algorithm.

Our scheduling method and policy have the potential to benefit all major players in the PCSH by improving patient outcomes and satisfaction, enabling providers to work on top of their license (i.e. better utilization of providers' time) and reducing operational inefficiencies. In this study, we focus on the impact of coordination on improving efficiencies and saving operational costs. Thus, we address the problem

that is one of the most perplexing to the medical community since most providers believe that patients and providers can benefit from better coordination of services. However, what is not clear is whether coordination of multiple services with its attendant increase in systems complexity will lead to higher cost of care.

Thus, we pose a series of research questions to be addressed, which are important to the management of the PCSH and other collaborative networks. First, is Silo a sustainable policy for the PCSH with same-day referrals? What is the opportunity cost if the PCSH continues with Silo? Second, how effective is the Static policy in coordinating APC and IMC patients, compared to Silo or a fully coordinated policy? Last, how much operational cost can the fully coordinated policy save for the PCSH? How much risk can it reduce, where the risk is measured as system variation? By comparing our coordinated scheduling method to other scheduling policies being considered by the PCSH, we show that the gain of switching to our proposed policy can be significant.

The contributions of this study are fourfold. First, to our knowledge, this is the first multi-server, multi-clinic model that studies sequential appointment scheduling in the literature. The model is based on a real healthcare network and it captures the essence of the uncertain, complicated patient flows in the network. Second, we propose an efficient scheduling method with a simple and effective booking limit to coordinate multiple clinics and to construct a balanced network schedule. Compared to other policies considered by the PCSH, our proposed policy yields high-quality and robust results. Third, we develop a novel scheduling algorithm that integrates analytical calculation, simulation, and statistical hypothesis testing to con-

struct a schedule. Lastly, we conduct a numerical study comparing our fully coordinated policy against Silo and Static. The results offer a fundamentally different way of structuring the PCSH to improve outpatient perioperative care that ensures close coordination between anesthesiologists and general internists. More generally, the results shed light on the risk in our increasingly interdependent healthcare system, if coordination is not properly implemented. Hence, we believe the insights are beneficial not only to the PCSH but other patient-centric models where scheduling coordination can be used.

The rest of this chapter is structured as follows. We briefly review the literature in Section 4.2. In Section 4.3, we provide our multi-server, multi-clinic model and our proposed scheduling policy. Section 4.4 explains the scheduling algorithm, with an emphasis on the hybrid evaluation method and the R&S procedure. In Section 4.5, we present our numerical study on various scheduling policies to address the aforementioned research questions and provide other insights. Section 4.6 concludes this chapter.

4.2. Related Literature

In previous chapters, we have reviewed the scheduling literature (Chapter 3) and presented studies related to the PCSH model (Chapter 2). So here, we limit ourselves to two streams of papers based on the two unique features in this study: a parallel server model and appointment scheduling using simulation.

We first review scheduling papers that employ a multi-server model. Note that

papers that assume a dedicated patient stream to each service provider are still single server models (Ahmadi-Javid et al. (2016)) and thus we do not include them in this review. Multi-server models are rich in the queueing literature, but, as pointed out by Erdogan and Denton (2013), the appointment scheduling problem differs from a typical queueing model in two main aspects. First, the steady state assumptions in queueing models do not hold in the transient clinical environment. Second, for most queueing models, it is assumed that patients arrive stochastically rather than according to a schedule. Liu and Liu (1998) consider a single clinic with multiple doctors whose arrival times are random. Given the number of patients to be scheduled, they use simulation to explore different schedules with different patient no-show types in each appointment block and find the best patient type mix and block assignment that minimizes doctor idleness and patient waiting. Sickinger and Kolisch (2009) model a single outpatient clinic with two CT scanners. Patients are screened by one of the two identical scanners that has a stochastic downtime. For a given number of patients, they propose a neighborhood search heuristic to find the best schedule that maximizes patient throughput. In a working paper by Zacharias and Pinedo (2016), they formulate a multi-server queueing model for a single clinic and study the impact of resource pooling on patient throughput and waiting times. Different from our study, they assume that the number of patients and their no-show types are known prior to scheduling. To the best of our knowledge, a network model with multiple servers in each station does not exist in the scheduling literature. We contribute to the literature by developing such a model and a scheduling methodology that is readily applicable to the real-world problem

in healthcare.

In the appointment scheduling literature, simulation has been used primarily as a tool to evaluate a solution, instead of generating one (Klassen and Yoogalingam (2009) and Ahmadi-Javid et al. (2016)). LaGanga and Lawrence (2007) study how to mitigate the detrimental impact of patient no-shows using appointment overbooking. They conduct simulation studies on various patient arrival patterns and establish the benefit of overbooking in increasing provider utility. Glowacka et al. (2009) develop a predictive model of patient no-shows and derive a set of sequencing rules to schedule patients offline. They use simulation to evaluate the performance of different sequencing rules and find the optimal number of patients to be scheduled. Klassen and Yoogalingam (2009) develop a heuristic to search for the optimal schedule and use simulation optimization to evaluate schedules offline. Different from the above studies, we use simulation to sequentially construct a schedule online. Moreover, we develop a hybrid method that expedites the simulation algorithm using analytical calculation and simulation. We contribute to the literature by developing an efficient scheduling method that integrates simulation procedures into sequential appointment scheduling.

4.3. Model Formulation

In this section, we formulate the PCSH model as a multi-server, multi-clinic healthcare network, where patients, after their nominal appointment, may be referred to another service on the same day. We leverage a similar slot model from Chapter 3. But different from Chapter 3, we develop a multiple server model where

providers work in parallel with stochastic service time and patients are seen by the first available provider in the clinic. We first present the model formulation and then illustrate how the assumption of parallel servers complicates the solution approach. In addition, we propose a myopic-based scheduling policy that coordinates different clinics to construct a balanced schedule that maximizes a network objective.

The PCSH consists of two clinics: APC and its supporting facility IMC (Figure 4.1). For notational purposes, we refer to APC and IMC as clinics 1 and 2, respectively. They operate during the same time period of the day, which is evenly divided into J time slots for appointment booking. There are n_1 and n_2 independent and identical providers at APC and IMC, respectively. Because the PCSH is implemented in a teaching hospital, providers (the servers), who are often medical residents or mid-level providers (e.g. nurse practitioners and physician assistants), do not have their own dedicated patients. Therefore, patients wait in one single queue at the beginning of each slot, and are seen by the first available provider in that clinic, first-come, first-serve. In queueing terms, providers are pooled resources that work in parallel. There are two types of patient requests, those who call for an appointment at APC (type 1) and those who call for IMC (type 2). The attendance rate for each request type is ρ_1 and ρ_2 for APC and IMC, respectively. The assumption of homogenous no-show within each clinic serves merely to simplify our mathematical expressions across this chapter. Our model and solution approach are readily applicable to heterogeneous patient no-show. To optimize the patients' overall health condition, APC patients are referred to IMC with probability $P_{1,2}$ for additional care, e.g. to control high blood pressure.

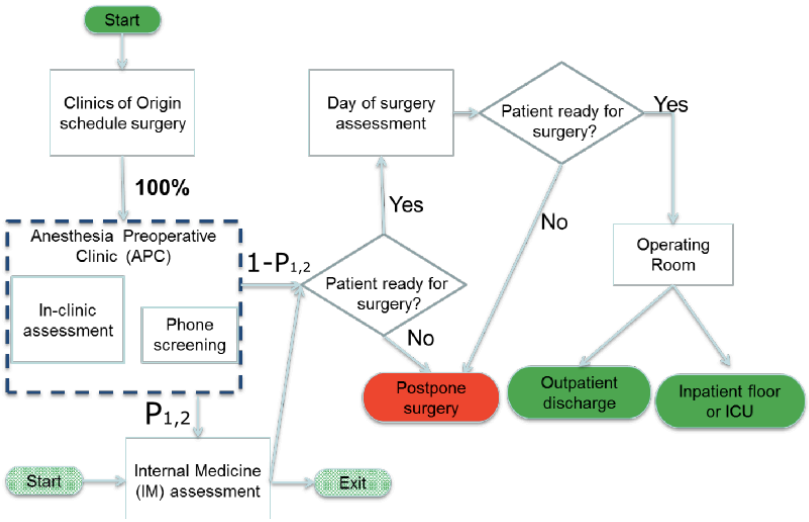

Figure 4.1: Simplified PCSH Model

To make an appointment, patients must to call the centralized scheduler in advance. Upon each request, the scheduler obtains the request type (APC or IMC), checks the current schedule and makes an assignment decision that maximizes the network objective function. We adopt a myopic objective, where each patient request is evaluated as the last request before the appointment day. The choice of a myopic approach instead of dynamic programming is based on: (1) the practicality of the model and solution approach, where DP is computationally intractable due to the high dimensionality of the problem (Lin et al. (2011)) and burdened with more modeling assumptions; (2) the fact that the myopic approach tends to yield good solutions (Chapter 3 and Muthuraman and Lawley (2008)).

Let $\mathbf{S}$ denote a network schedule and $S_{i,j}$ be the number of scheduled patients in clinic i slot j, abbreviated as (i, j). Due to no-show, the arrival variable, $X_{i,j}$, follows

a binomial distribution

$$P(X_{i,j} = m \mid S_{i,j}, \rho_i) = \begin{pmatrix} S_{i,j} \\ m \end{pmatrix} \rho_i^m \cdot (1-\rho_i)^{S_{i,j}-m}. \tag{4.1}$$

At each clinic, the service time per patient is assumed to be independent and identically distributed and follow the exponential distribution. The choice of exponential service time results from the trade-off between analytical tractability and the practicality of highly uncertain service times that cannot be approximated by assuming deterministic service times. The latter follows from the fact that the PCSH tends to see more complicated heterogeneous patients that need to be optimized for surgery (Ferschl et al. (2005)). Let λ_i be the service rate of each provider at clinic i. Let $L_{i,j}$ be the capacity of a provider at (i,j), which is defined as the number of services this provider is able to complete if unlimited number of patients are waiting for service. Since service times are exponential, we have $L_{i,j} \sim \text{Poisson}(\lambda_i)$. The number of completed services from (i,j), $Z_{i,j}$, is a random variable that depends on the pooled resource capacity at (i,j) and the number of patients waiting for service at the beginning of (i,j), denoted as $A_{i,j}$. After the service, some of the APC patients are referred to IMC for additional care. Under the PCSH, APC referrals are guaranteed for immediate access to IMC, so they can join the queue of the immediate next slot at IMC. The number of referral patients leaving $(1,j)$ for $(2,j+1)$ is denoted as $R_{1,j}$, which when conditioned on $Z_{1,j}$ follows a binomial distribution

$$P(R_{1,j} = r | Z_{1,j}) = \begin{pmatrix} Z_{1,j} \\ r \end{pmatrix} (P_{1,2})^r \cdot (1-P_{1,2})^{Z_{1,j}-r}. \tag{4.2}$$

Patients who do not complete their services overflow to the next time slot and join the queue of patients at $(i,j+1)$. Let the number of overflows from (i,j) to $(i,j+1)$

be denoted as $Y_{i,j}$. By the balance of flow, we have, for $j = 1, 2, ..., J$,

$$A_{1,j} = Y_{1,j-1} + X_{1,j} = Z_{1,j} + Y_{1,j},$$

$$A_{2,j} = Y_{2,j-1} + X_{2,j} + R_{1,j-1} = Z_{2,j} + Y_{2,j}, \tag{4.3}$$

where $Y_{i,0} = 0$. As shown in the above equations, the queue at the beginning of an APC slot is made up of exogenous type 1 arrivals and overflows from the previous slot. Additionally, the queue at IMC consists of exogenous type 2 arrivals, overflows from the previous slot and APC referrals.

Under the myopic policy, the objective of the scheduler is to assign each patient request to a slot that maximizes

$$V(\mathbf{S}^n) = E(Revenue(\mathbf{S}^n)) - E(Cost(\mathbf{S}^n)) \tag{4.4}$$

where $\mathbf{S}^n$ is the resulting schedule from a slotting decision and the superscript n denotes the number of patients in the schedule. The network is rewarded for each service provided to APC and IMC patients, r_1 and r_2, respectively. Because APC patients might generate additional reward from referral service, the expected bundled reward of an APC patient is $r_1 + r_2 P_{1,2}$. Therefore, the total reward from $\mathbf{S}^n$ is

$$E(Revenue(\mathbf{S}^n)) = \sum_j \left(S^n_{1,j} \cdot \rho_1 \cdot (r_1 + P_{1,2} \cdot r_2) + S^n_{2,j} \cdot \rho_2 \cdot r_2 \right). \tag{4.5}$$

To prevent excessive patient waiting and clinical overtime, we penalize each occurrence of patient overflow by $c_{i,j}$, depending on the clinic and slot. When patients overflow from slot J, providers have to work overtime, and the cost penalty

is usually higher than the reward, i.e., $c_{1,J} > r_1 + r_2 P_{1,2}$ and $c_{2,J} > r_2$. Therefore, the expected cost from $\mathbf{S}^n$ is

$$E(Cost(\mathbf{S})) = \sum_j P(Y_{1,j} = y_{1,j},\, Y_{2,j} =_{2,j}) \cdot (y_{1,j}\, y_{2,j}) \cdot \begin{pmatrix} c_{1,j} \\ c_{2,j} \end{pmatrix}. \tag{4.6}$$

As can be seen from Equation (4.6), the complexity in evaluating the objective function resides in the joint overflow probability distribution, $P(Y_{1,j},\, Y_{2,j})$, which is derived in the next subsection.

4.3.1 Derivation of the Joint Overflow Distribution

From Equation (4.3), we know that the joint overflow distribution at slot j, $P(Y_{1,j}, Y_{2,j})$, depends on four variables: completed services, $\mathbf{Z}_{.,j}$, exogenous arrivals, $\mathbf{X}_{.,j}$, the conditional referral and overflows from an earlier slot, $R_{1.,j-1}$ and $\mathbf{Y}_{.,j-1}$, respectively. The recursive dependency of the overflow variables, together with the referral variable that intertwines the two clinics, drives the high dimensionality of this model.

We start with a sketch of how the overflow variable relates to the aforementioned patient flow variables. Then, we derive the distribution of each of these variables. Lastly, we combine the results and present the distribution of the joint overflow variable in Proposition 4.3.3. To simplify the notation, we omit the conditional term, $|\mathbf{S}$, from all the expressions in the following paragraphs. For example, $P(\mathbf{Y}_{.,\mathbf{j}}|\mathbf{S})$ is abbreviated as $P(\mathbf{Y}_{.,\mathbf{j}})$.

To derive $P(\mathbf{Y}_{.,j} = \mathbf{y}_{.,j})$, we first condition on the number of patients waiting at

the beginning of slot j,

$$P(\mathbf{Y}_{\cdot,j} = \mathbf{y}_{\cdot,j}) = \sum_{\mathbf{a}_{\cdot,j} \in \Omega_a^j} P(\mathbf{Y}_{\cdot,j} = \mathbf{y}_{\cdot,j} | \mathbf{A}_{\cdot,j} = \mathbf{a}_{\cdot,j}) P(\mathbf{A}_{\cdot,j} = \mathbf{a}_{\cdot,j}), \tag{4.7}$$

where the support set of $\mathbf{a}_{\cdot,j}$, Ω_a^j, is defined as

$$\Omega_a^j := \left\{ \times_{i \in \mathbf{I}} \left\{ a_{i,j} \in \mathbb{Z} : \, 0 \leq a_{i,j} \leq U(i,\, j) \right\} \right\} \bigcap \left\{ \sum_{i \in \mathbf{I}} a_{i,j} \leq \sum_{i \in \mathbf{I}} \sum_{j' \leq j} S_{i,j'} \right\},$$

and $U(i,j) \equiv \sum_{i \in \mathbf{I}} \sum_{j' < j} S_{i,j'} + S_{i,j}$, which is the maximum number of patients waiting at (i,j). By Equation (3.2), we have $\mathbf{A}_{\cdot,j} = \mathbf{Z}_{\cdot,j} + \mathbf{Y}_{\cdot,j}$ and we can express the first term in Equation (4.7) with $\mathbf{Z}_{\cdot,j} | \mathbf{A}_{\cdot,j}$. Moreover, conditioning on $\mathbf{A}_{\cdot,j}$, the completed service variables across different clinics are independent, so $P(\mathbf{Z}_{\cdot,j} = \mathbf{z}_{\cdot,j} | \mathbf{A}_{\cdot,j} = \mathbf{a}_{\cdot,j}) = \prod_i P(Z_{i,j} = z_{i,j} \, | A_{i,j} = a_{i,j})$. Proposition 4.3.1 specifies the distribution of $P(Z_{i,j} | A_{i,j})$. The distribution of $P(\mathbf{A}_{\cdot,j} = \mathbf{a}_{\cdot,j})$ is derived in Proposition 4.3.2.

Proposition 4.3.1. *Under exponential service time, the completed service distribution, $P(Z_{i,j} = z \, | A_{i,j} = a)$, can be expressed as*

$$P(Z_{i,j} = z | A_{i,j} = a) = \left(\begin{array}{c} a \\ z \end{array} \right) P(L_{i,j} = 0)^{a-z} \cdot P(L_{i,j} \geq 1)^z,$$

for $a \leq n_i$, and for $a > n_i$,

$$\begin{aligned} P(Z_{i,j} = z | A_{i,j} = a) =& 1(a - z > n_i) \cdot \sum_{\{z^1, z^2, \ldots, z^{n_i}\} \in \Omega_{n_i}^z} \prod_{k=1}^{n_i} P(L_{i,j} = z^k) \\ &+ 1(a - z \leq n_i) \cdot \left(\begin{array}{c} n_i \\ a - z \end{array} \right) \\ &\cdot \sum_{z_b = 0}^{a - n_i} \Big(\sum_{\{z^1, z^2, .., z^{a-z}\} \in \Omega_{a-z}^{z_b}} \prod_{k=1}^{a-z} P(L_{i,j} = z^k) \Big) \\ &\cdot \Big(\sum_{\{z^1 \geq 1, z^2 \geq 1, .., z^{n_i - a + z} \geq 1\} \in \Omega_{n_i - a + z}^{z - z_b}} \prod_{k=1}^{n_i - a + z} P(L_{i,j} \geq z^k) \Big), \end{aligned}$$

where $\Omega_n^z = \{z^k \geq 0,\, z^k \in \mathbb{Z},\, k = 1, 2, ..., n : z^1 + z^2 + ... + z^n = z\}$.

Proof. When $a \leq n_i$, the number of patients waiting at the beginning of slot j is smaller than the number of providers at clinic i. Therefore, all a patients start their services immediately, the queue empties and there are $(n_i - a)$ providers unutilized. By the end of slot j, z patients completes their services, which implies that (1) the providers of these z patients each has the capability to see at least one patient, i.e. $P(L_{i,j} \geq 1)$; and (2) the other $(a - z)$ providers who do not finish their services each has a capacity of less than one, i.e. $P(L_{i,j} = 0)$. The term $\begin{pmatrix} a \\ z \end{pmatrix}$ lists all combinations that z patients among a total of a patients complete their services. We remind the reader that $L_{i,j}$ is the capacity of each provider. Since providers are assumed identical and independent, we can omit a specific provider subscript on $L_{i,j}$ and write the joint probability as a product of the individual $L_{i,j}$ probabilities. As a result, when there are fewer patients than the providers, $P(Z_{i,j} = z | A_{i,j} = a)$ is binomial distribution with parameters $(a,\, P(L_{i,j} \geq 1))$.

When $a > n_i$, there are more patients than providers and thus some patients have to wait at the beginning of slot j. If $z = 0$, no provider completes his or her initial service, and $L_{i,j} = 0$ for all n_i providers. The second equation in Proposition 4.3.1 simplifies to

$$P(Z_{i,j} = 0 | A_{i,j} = a) = P(L_{i,j} = 0)^{n_i}.$$

For $z \geq 1$, we need to consider two scenarios: whether or not the patient queue empties by the end of slot j. The first indicator function in the second equation of Proposition 4.3.1, $1(a - z > n_i)$, represents the case where the queue is non-empty.

The second indicator function, $1(a - z \leq n_i)$, represents the empty queue case. In the first scenario, the pooled capacity of the n_i providers equals z. Let z^k be the number of services completed by the kth provider. We define Ω_n^z to be the set of the number of services each provider completes so that the total number of completed services is z. That is,

$$\Omega_n^z = \{z^k \geq 0,\, z^k \in \mathbb{Z},\, k = 1, 2, ..., n : z^1 + z^2 + ... + z^n = z\}.$$

Therefore, if $a - z > n_i$, the probability z out of a patients complete their service by the end of the slot is

$$P(Z_{i,j} = z | A_{i,j} = a,\, a - z > n_i) = \sum_{\{z^1, z^2, ..., z^{n_i}\} \in \Omega_{n_i}^z} \prod_{k=1}^{n_i} P(L_{i,j} = z^k). \tag{4.8}$$

For the scenario where the patient queue empties at the end of slot j, we know that the uncompleted services are each with a provider. So $(a - z)$ providers are still busy with patients and the rest $(n_i - a + z)$ providers are idle. The term $\begin{pmatrix} n_i \\ a - z \end{pmatrix}$ enumerates all cases that $(a - z)$ out of n_i providers are busy. Then, for a particular enumeration, we condition on the number of services the $(a - z)$ busy providers have jointly completed, denoted as z_b. Hence, $(z - z_b)$ is the number of services jointly completed by the idle providers. The value z_b is between zero and $(a - n_i)$. The upper bound is because each idle provider must have completed his or her first

service, so $(z - z_b) \geq n_i - a + z$. Therefore, when $a - z \leq n_i$, we have

$$P(Z_{i,j} = z | A_{i,j} = a,\, a - z \leq n_i) = \begin{pmatrix} n_i \\ a - z \end{pmatrix} \sum_{z_b=0}^{a-n_i} \Big(\sum_{\{z^1, z^2, .., z^{a-z}\} \in \Omega_{a-z}^{z_b}} \prod_{k=1}^{a-z} P(L_{i,j} = z^k) \cdot \sum_{\{z^1 \geq 1, z^2 \geq 1, .., z^{n_i - a + z} \geq 1\} \in \Omega_{n_i - a + z}^{z - z_b}} \prod_{k=1}^{n_i - a + z} P(L_{i,j} \geq z^k) \Big) \tag{4.9}$$

The first term in Equation (4.9) after the summation over z_b is the probability that the $(a - z)$ busy providers complete z_b services, which is their pooled capacity. The second term is the probability that the idle providers complete $(z - z_b)$ services, which is the lower bound on their pooled capacity, because there is no more patients to be seen. In other words, some of these idle providers may be able to see more patients if there were patients waiting in the queue. Note that the value of z^k for idle providers is lower bounded by 1, because being idle implies that the provider has at least completed his or her initial service.

Finally, combining the two scenarios, Equations (4.8) and (4.9), yields the second equation in Proposition 4.3.1. □

Proposition 4.3.2. *The distribution of total number of patients waiting at j, $P(\mathbf{A}_{\cdot,j} = \mathbf{a}_{\cdot,j})$, is*

$$P(\mathbf{A}_{\cdot,1} = \mathbf{a}_{\cdot,1}) = \textstyle\prod_i P(X_{i,1} = a_{i,1})$$

for $j = 1$, and for $j > 1$,

$$P(\mathbf{A}_{.,j} = \mathbf{a}_{.,j}) = \sum_{0 \le x_{i,j} \le S_{i,j}, \forall i} \prod_i P(X_{i,j} = x_{i,j}) \Bigg(\sum_{\mathbf{a}_{.,j-1} \in \Omega_a^{j-1}} P(\mathbf{A}_{.,j-1} = \mathbf{a}_{.,j-1}) \cdot \Big(\sum_{\mathbf{z}_{.,j-1} \in \Omega_z^{j-1}} P(Z_{1,j-1} = a_{1,j-1} - a_{1,j} + x_{1,j}, Z_{2,j-1} = z_{2,j-1} | \mathbf{a}_{.,j-1}) \cdot P(R_{1,j-1} = a_{2,j} - x_{2,j} - a_{2,j-1} + z_{2,j-1} | Z_{2,j-1} = z_{2,j-1}) \Big) \Bigg).$$

Proof. When $j = 1$, $\mathbf{A}_{.,1} = \mathbf{X}_{.,1}$ and due to the independence of exogenous arrivals, $P(\mathbf{X}_{.,1}) = \prod_i P(X_{i,j})$, whose distribution has been shown in Equation (4.1). For $j > 1$, $A_{1,j} = X_{1,j} + Y_{1.,j-1}$ and $A_{2,j} = X_{2,j} + Y_{2,j-1} + R_{1,j-1}$. Because of the referral variable, $A_{1,j}$ and $A_{2,j}$ are interdependent and need to be expressed jointly. We first condition on the exogenous arrivals and express $\mathbf{A}_{.,j}$ in terms of the patient flow variables,

$$P(\mathbf{A}_{.,j} = \mathbf{a}_{.,j}) = \sum_{0 \le x_{1,i} \le S_{1,i}, \forall i} \prod_i P(X_{i,j} = x_{i,j}) \cdot P(Y_{1,j-1} = a_{1,j} - x_{1,j},\ Y_{2,j-1} + R_{1,j-1} = a_{2,j} - x_{2,j}). \tag{4.10}$$

Note that the conditional term on $\mathbf{X}_{.,j}$ is dropped from the second term, because of independence. Now, we condition the second term on $\mathbf{A}_{.,j-1}$ and have

$$\begin{aligned} &P(Y_{1,j-1} = a_{1,j} - x_{1,j},\ Y_{2,j-1} + R_{1,j-1} = a_{2,j} - x_{2,j}) \\ &\quad = \sum_{\mathbf{a}_{.,j-1} \in \Omega_a^{j-1}} P(\mathbf{A}_{.,j-1} = \mathbf{a}_{.,j-1}) \\ &\qquad P(Y_{1,j-1} = a_{1,j} - x_{1,j},\ Y_{2,j-1} + R_{1,j-1} = a_{2,j} - x_{2,j} | \mathbf{A}_{.,j-1} = \mathbf{a}_{.,j-1}). \end{aligned} \tag{4.11}$$

Note that, we now observe the recursive structure that $P(\mathbf{A}_{.,j})$ relies on $P(\mathbf{A}_{.,j-1})$. In the second term of the above equation, we replace $Y_{i,j-1}$ by the corresponding

$Z_{i,j-1} = A_{i,j-1} - Y_{i,j-1}$ and obtain

$$\begin{aligned}
&P(Y_{1,j-1} = a_{1,j} - x_{1,j},\, Y_{2,j-1} + R_{1,j-1} = a_{2,j} - x_{2,j} | \mathbf{A}_{.,j-1}) \\
&\quad = P(a_{1,j-1} - Z_{1,j-1} = a_{1,j} - x_{1,j},\, a_{2,j-1} - Z_{2,j-1} + R_{1,j-1} = a_{2,j} - x_{2,j} | \mathbf{a}_{.,j-1}) \\
&\quad = \sum_{\mathbf{z}_{.,j-1} \in \Omega_z^{j-1}} P(Z_{1,j-1} = a_{1,j-1} - a_{1,j} + x_{1,j},\, Z_{2,j-1} = z_{2,j-1} | \mathbf{a}_{.,j-1}) \cdot \\
&\qquad P(R_{1,j-1} = a_{2,j} - x_{2,j} - a_{2,j-1} + z_{2,j-1} | Z_{2,j-1} = z_{2,j-1}),
\end{aligned} \tag{4.12}$$

where the support set for $\mathbf{z}_{.,j-1}$, Ω_z^{j-1}, is defined as the following,

$$\Omega_z^{j-1} := \left\{ \times_{i \in \mathbf{I}} \left\{ z_{i,j-1} \in \mathbb{Z} :\, 0 \leq z_{i,j-1} \leq \sum_{i \in \mathbf{I}} U_{i,j-1} \right\} \right\} \bigcap \left\{ \sum_{i \in \mathbf{I}} z_{i,j-1} \leq \sum_{i \in \mathbf{I}} \sum_{j' \leq j-1} S_{i,j'} \right\}.$$

The first term limits the maximum value individual $z_{i,j-1}$ can take. The second term ensures that the joint value of completed services is feasible. Substituting the results in Proposition 4.3.1 and Equation (4.2) into Equation (4.12) completes the derivation of $P(Y_{1,j-1},\, Y_{2,j-1} + R_{1,j-1} | \mathbf{A}_{.,j-1})$. Then, combining Equations (4.10), (4.11) and (4.12) yields Proposition 4.3.2 for $j > 1$. □

Lastly, substituting Propositions 4.3.1 and 4.3.2 into Equation (4.7) yields the following proposition.

Proposition 4.3.3. *The joint overflow distribution, $P(\mathbf{Y}_{.,j} = \mathbf{y}_{.,j})$, is*

$$P(\mathbf{Y}_{.,j} = \mathbf{y}_{.,j}) = \prod_{i=1}^{2} \Big(\sum_{x_{i,1}=0}^{S_{i,1}} P(X_{i,1} = x_{i,1}) P(Z_{i,1} = x_{i,1} - y_{i,1} | A_{i,1} = x_{i,1}) \Big),$$

for $j = 1$ and for $j > 1$, it is

$$P(\mathbf{Y}_{.,\mathbf{j}} = \mathbf{y}_{.,\mathbf{j}}) = \sum_{\mathbf{a}_{.,\mathbf{j}} \in \Omega_a^j} \left(P(\mathbf{A}_{.,j} = \mathbf{a}_{.,j}) \cdot \prod_{i=1}^{2} \Big(\binom{a}{z} P(L_{i,j} = 0)^{a-z} \cdot P(L_{i,j} \geq 1) \Big) \right),$$

when $a \leq n_i$, *and*

$$
\begin{aligned}
P(\mathbf{Y}_{.,\mathbf{j}} = \mathbf{y}_{.,\mathbf{j}}) \quad = \quad & \sum_{\mathbf{a}_{.,\mathbf{j}} \in \Omega_a^j} P(\mathbf{A}_{.,j} = \mathbf{a}_{.,j}) \cdot \prod_{i=1}^{2} \Bigg(1(a - z > n_i) \\
& \cdot \sum_{\{z^1, z^2, \ldots, z^{n_i}\} \in \Omega_{n_i}^z} \prod_{k=1}^{n_i} P(L_{i,j} = z^k) \\
& + 1(a - z \leq n_i) \cdot \begin{pmatrix} n_i \\ a - z \end{pmatrix} \\
& \cdot \sum_{z_b=0}^{a-n_i} \Big(\sum_{\{z^1, z^2, .., z^{a-z}\} \in \Omega_{a-z}^{z_b}} \prod_{k=1}^{a-z} P(L_{i,j} = z^k) \Big) \\
& \cdot \Big(\sum_{\{z^1 \geq 1, z^2 \geq 1, .., z^{n_i - a + z} \geq 1\} \in \Omega_{n_i - a + z}^{z - z_b}} \prod_{k=1}^{n_i - a + z} P(L_{i,j} \geq z^k) \Big) \Bigg)
\end{aligned}
$$

when $a > n_i$.

4.3.2 Restricted Coordinated Myopic Policy (RC Policy)

Due to the challenge of the high complexity of patient flow dynamics, we adopt a myopic-based solution approach. The major concern with using a myopic approach is the shortsightedness that may result in less satisfying solutions due to a "bad" call-in sequence. In the case of the PCSH, such suboptimal schedules occur when the early requests are dominated by exogenous IMC requests, which blocks referral flows from APC and limits the number of APC patients that can be scheduled. This contradicts the PCSH goal of guaranteeing APC patients same day referrals. However, in sequential scheduling, it is very challenging to foresee the optimal number of patients in each clinic prior to scheduling. Therefore, we propose a restricted coordinated policy (RC in short) that mitigates the issue of being myopic.

To prevent locking into a suboptimal schedule due to unfavorable call-in sequences, we impose a simple booking limit on the IMC requests.

The idea of the RC policy is to preserve capacity at IMC in anticipation of APC referrals. The amount of capacity reserved is proportional to the expected number of referrals and is calculated in the following manner. We first define the capacity of a clinic i as the product of its service rate, the number of servers and the slots: $\lambda_i \cdot n_i \cdot J$. It is independent of its patients' no-show rate because the scheduler will try to match clinical capacity with actual demand via overbooking. Given the capacity of APC, the expected number of referrals is $\lambda_1 \cdot n_1 \cdot J \cdot P_{1,2}$, which is the amount of IMC resources needed by referrals. Because an APC referral arrives at the immediate subsequent slot after its service, the first slot in IMC will not be used for referral patients. Also, referrals from slot J are seen during IMC overtime. Therefore, the capacity left for exogenous IMC patients is $\max(\lambda_2, \lambda_2 n_2 J - \lambda_1 n_1 J P_{1,2})$. Converting the leftover capacity to the number of exogenous IMC patients yields the booking limit (BL):

$$BL = \frac{\max(\lambda_2 n_2, \lambda_2 n_2 J - \lambda_1 n_1 J P_{1,2})}{\rho_2}. \tag{4.13}$$

The booking limit is a simple and effective modification of the unrestricted coordinated myopic policy developed in Chapter 3 (the UC policy, for short). It is not burdened by a heavily parameterized model on the call-in process as required in DP, but provides an effective measure to balance the utilization of both APC and IMC in the PCSH. In addition, by restricting the number of patients instead of blocking certain slots in IMC, the scheduler can dynamically balance the utilization in

both clinics during the call-in process. Zeng et al. (2009) propose a similar strategy in their single-clinic scheduling when dealing with heterogenous no-shows. They restrict the number of patients with high no-shows to 4 or 8 and show the improvement against the myopic policy without restriction. However, they do not specify how to find a booking limit on high no-show patients beyond trial and error.

4.4. Scheduling Algorithm

In this section, we introduce the scheduling algorithm under the RC policy that sequentially constructs a network schedule based on patient requests. The algorithm consists of two critical components: the evaluation of candidate schedules, and selecting the best schedule. Theoretically, we can calculate the expected profit of any schedule using Equation (4.4). However, the high dimensionality in our multi-clinic, multi-server model prohibits analytical calculation for the state space of all possible schedules, even for a moderate-sized network. To overcome this challenge, we develop a hybrid method that analytically calculates the expected profit of schedules that are relatively empty, and simulates the profits as schedules become fuller. Due to simulation randomness, we employ an R&S procedure to select the best schedule, based on the Kim-Nelson (KN) procedure (Kim and Nelson (2001)). R&S is a hypothesis testing method for multiple comparisons. Given the predetermined parameters α and δ, R&S guarantees with $(1-\alpha)$ percent confidence that the selected schedule has an expected profit that is within δ units of the schedule with the highest expected profit.

We first define the parameters and variables in Table 4.1 and then explain the

scheduling algorithm in Algorithm 4.1. The details of the hybrid evaluation method and the R&S procedure in Steps 3 and 5 of Algorithm 4.1 are presented and explained in Subsections 4.4.1 and 4.4.2, respectively.

Parameters	Definition
BL_i	booking limit for clinic i
$\mathbf{S}^{th}$	threshold schedule to switch the evaluation method from analytical to simulation
α	significance level for R&S
δ	indifference value for R&S
N_o	number of simulation batches in each sample
N_b	batch size, and $N_o \cdot N_b$ is simulation sample size
$\Delta_{i,j}$	an $I \times J$ matrix with value 1 in cell (i, j) and 0 elsewhere.

Variables	Definition	Initial Value
n	the number of patients in schedule	0
$\mathbf{S}^n$	the current schedule	$\mathbf{S}^0$ is an $I \times J$ null matrix
Ψ	a set of rejected request types	$\Psi = \emptyset$
Set_J	a set of candidate slots	$\{0, 1, 2, ..., J\}$, 0 is rejection
Set_C	a set of candidate schedules that correspond to the candidate slots in Set_J	$\emptyset$
Set_A	a set of slots whose schedules' expected profits are analytically calculated	$\emptyset$
Set_S	a set of slots whose schedules' expected profits are simulated	$\emptyset$
$b_{R\&S}$	a binary variable to track if R&S is used	1 if R&S is used and 0 otherwise
N_s	total number of batches	0
$Q_{k,j}$	the k^{th} observation of slot j	
$\overline{\mathbf{Q}_j}(m)$	the sample average of the first m observations at slot j	
D	indifference value of the final schedule	0
CL	confidence level of the final schedule	100%

Table 4.1: Notations

Algorithm 4.1 Scheduling Algorithm

Step 0. Initialize all parameters and variables in Table 4.1.
Step 1. Wait for the next patient request.
Initialize $b_{R\&S}$, N_s, Set_J, Set_A, Set_S, Set_C.
Step 2. A patient calls and request clinic i.
If $i \in \Psi$, reject and go to Step 1.
Elseif $\sum_j \mathbf{S}^n_{i,j} = BL_i$, reject the request, include i in Ψ and go to Step 1.
Else, $\forall j \in Set_J$, set $Set_C = Set_C \cup \{\mathbf{S}^n + \Delta_{i,j}\}$, and go to Step 3.
Step 3. **Hybrid Evaluation Method**
Use Algorithm 4.2 to calculate the expected profit for each candidate schedule in Set_C.
Step 4. If Set_C contains more than one schedule, go to Step 5.
Otherwise, go to Step 6.
Step 5. **Ranking and Selection Procedure**
Apply R&S procedure in Algorithm 4.3 over Set_C and Set_J.
Set $b_{R\&S} = 1$.
If more than one schedule remains in Set_C, go to Step 3.
Otherwise, go to Step 6.
Step 6. **Decision**
Set $D = D + b_{R\&S} \cdot \delta$, $CL = CL \cdot (1-\alpha)^{1-b_{R\&S}}$.
Offer $j \in Set_J$ to the patient.
If $j \neq 0$, set $\mathbf{S}^{n+1} = \mathbf{S}^n + \Delta_{i,j}$ and $n = n+1$.
Otherwise, include i in Ψ.
Step 7. **Stopping Rule**
If $\Psi = I$, terminate the algorithm. Otherwise, go to Step 1.

Referring to Algorithm 4.1, the scheduler starts with an empty schedule. When a patient calls with clinic request, the scheduler checks if the requested clinic is on the rejection list Ψ, or it has reached its booking limit. If so, the request is rejected and may be considered for another appointment day. Otherwise, the scheduler updates the set of candidate schedules (Set_C). Because there is an one-to-one correspondence between the elements of Set_C and Set_J, for simplicity, we explain the algorithm in terms of the candidate schedules and Set_C. The scheduler then evaluates the expected profits of candidate schedules, using the hybrid evaluation method de-

scribed in Subsection 4.4.1. Given the profit data of all candidate schedules in Set_C, a R&S procedure is employed to select the best schedule with high confidence, if there are multiple candidates in Set_C. If more data is needed to select such a schedule, the algorithm repeats Steps 3 to 5, until a single schedule remains in Set_C. The the corresponding slot ($j \in Set_J$) is offered to the patient and the schedule is updated. If the decision is to reject, that is, adding the patient decreases the profit, the request type is recorded in Ψ and later requests of the same type will be automatically rejected. Again, rejected patients may be considered for another appointment day. The variables CL and D are updated after each patient request. The algorithm terminates when all requests types are included in Ψ, otherwise it returns to Step 1 to wait for the next patient request. When the algorithm terminates, we have constructed the final schedule for the PCSH, with a confidence level of at least CL and within D units of expected profit of the true best schedule.

4.4.1 Hybrid Evaluation Method (Step 3)

Algorithm 4.2 provides details for the hybrid evaluation method. The analytical calculation using Equation (4.4) and Propositions 4.3.3 has two major advantages. First, the results can be re-used. Second, the schedule with the highest expected profit is the best. However, the analytical calculation becomes computationally infeasible as the schedule fills up with more patients. Simulation, on the other hand, enables complicated computations in a reasonable amount of time but introduces uncertainty in its solution. As a result of simulation variance, the scheduler can no longer select a schedule simply based on its expected profit.

Algorithm 4.2 Hybrid Evaluation Method

Step 1. Set $N_s = N_s + N_o$.
Step 2. For every $j \in Set_J$,
  if $\mathbf{S}^n + \Delta_{i,j} \leq \mathbf{S}^{th}$
    compute $V(\mathbf{S}^n + \Delta_{i,j})$, duplicate its value N_o times
    store them in $Q_{k,j}$, for $k = (N_s - N_o + 1)$, $(N_s - N_o + 2)$,..., N_s
    and include j in Set_A.
  otherwise,
    simulate $V(\mathbf{S}^n + \Delta_{i,j})$ over $N_0 \cdot N_b$ replications,
    batch data into N_o batches of size of N_b,
    store batch averages in $Q_{k,j}$, for $k = (N_s - N_o + 1)$, $(N_s - N_o + 2)$,..., N_s
    and include j in Set_S.
Step 3. If Set_A is non-empty,
  let $j^* = \arg\max_{j \in Set_A}\{V(\mathbf{S}^n + \Delta_{i,j})\}$ and $Set_J = Set_S \cup \{j^*\}$.
  for every $j \neq j^*$, $j \in Set_A$, remove $(\mathbf{S}^n + \Delta_{i,j})$ from Set_C.
Otherwise, set $Set_J = Set_S$.

To leverage on the advantages of both methods, we define a threshold schedule, $\mathbf{S}^{th}$, that determines when to switch from analytical calculation to simulation, as a schedule fills up. One can choose such an $\mathbf{S}^{th}$ based on the computational power or time allowance. In our study, we set $\mathbf{S}^{th}$ to be the schedule where each provider has one patient. For example, for an I-clinic, J-slot network with n_i providers at clinic i, $\mathbf{S}^{th}(i,j) = n_i$. Based on experimentation, we found that this choice of $\mathbf{S}^{th}$ provided the best performance for our algorithm. Under our hybrid evaluation method, Step 3 in Algorithm 4.1 is evaluated analytically, when $\mathbf{S}^n(i,j) \leq \mathbf{S}^{th}(i,j), \forall i,j$, and is approximated using simulation otherwise.

The hybrid evaluation method in Algorithm 4.2 starts by updating the variable N_s, which tracks the total number of batches that will be sampled after Algorithm 4.2. In Step 2, each candidate schedule in Set_C is evaluated, either analytically or using simulation. For schedules whose profits are analytically calculated, we first

duplicate its value N_o times (to prepare for R&S), assign them to the corresponding variables, and include the corresponding slots in Set_A. For the schedules whose profits are simulated, their corresponding slots are included in Set_S. To simulate the expected profit of a schedule, we sample $N_o \cdot N_b$ replications, batch the data into N_o batches of size N_b, and assign the batch averages to the corresponding variables. The reason for batching is to comply to the normality assumption required in R&S. Lastly, we update Set_C in Step 3 to prepare for R&S. Only the best analytical schedule (if Set_A is non-empty) is included in Set_C, as well as all simulated schedules. We update Set_J accordingly.

4.4.2 Ranking and Selection Procedure (Step 5)

The R&S procedure is a statistical hypothesis testing method that screens out inferior decisions given a confidence level and indifference value. We employ the R&S procedure described in Kim and Nelson (2001). To fit in our scheduling algorithm (Algorithm 4.1), the implementation of R&S differs from that in KS in three ways. First, because of our hybrid evaluation method, not all data is simulated, and screening is only applied to a subset of candidate schedules instead of the entire set as in KN. To be efficient, we separate data generating (Step 3 in Algorithm 4.1) from R&S (Step 5 in Algorithm 4.1). Second, to optimize parallel computing, when more data is needed, we sample another N_0 observations, while KN resamples one observation a time. Third, in the context of sequential scheduling, the interpretation of the confidence level and the indifference value in our final schedule is different from NS. We apply R&S to each patient request and the probability of correct selection is

at least $(1-\alpha)$, conditioning on the current state of the schedule. So the probability of constructing the best final schedule is the probability of all correct slotting decisions for every patient request, which is lower bounded by $(1-\alpha)^n$, where n is the number decisions made by R&S. The quantity $(1-\alpha)^n$ is a lower bound because $(1-\alpha)$ is the lower bound confidence level for each patient request that employs R&S. In terms of the indifference value, the maximum deviation our final schedule from the best is no more than $n \cdot \delta$. The reason is that, at each patient request, the schedule selected by R&S is at most δ units lower (in expected profit) than the best achievable schedule.

Algorithm 4.3 Ranking and Selection Procedure (modified from KN)

Step 0. Obtain parameters and data from Algorithm 4.1

$\mathbf{S}^n$, α, δ, N_o, Set_J, Set_C, N_s

Compute h^2 using the following equation,

$$h^2 = \left[\left(\frac{2\alpha}{k-1}\right)^{-2/(N_0-1)} - 1\right] \cdot (N_0 - 1)$$

where k is the number of slots in Set_J.

Step 1. **Variance of the difference between two schedules**

For $j \neq l$ and $j, l \in Set_J$, compute

$$\Gamma_{j,l} = \frac{1}{N_0-1} \sum_{n=1}^{N_o} \left(Q_{n,j} - Q_{n,l} - \left(\overline{\mathbf{Q}}_j(N_o) - \overline{\mathbf{Q}}_l(N_o)\right)\right)^2$$

$$N_j = \max_{l \neq j} \left\{ \left\lfloor \frac{h^2 \Gamma_{j,l}}{\delta^2} \right\rfloor \right\}$$

where $\lfloor . \rfloor$ indicates the integer part of the expression value.

Step 2. **Selection by sufficient sample size**

If $N_s > \max_{j \in Set_J}\{N_j\}$,

set $j^* = \arg\max_j \{\overline{\mathbf{Q}}_j(N_s)\}$, set $Set_J = j^*$, $Set_C = \mathbf{S}^n + \Delta_{i,j^*}$,

and terminate the R&S procedure.

Otherwise, go to Step 3.

Step 3. **Screen out inferior schedules**

Set $oldSet_J = Set_J$.

For $j \neq l$ and $j, l \in oldSet_J$, compute

$$W_{j,l}(N_s) = \max\left\{0, \frac{\delta}{2N_s} \cdot \left(\frac{h^2 \Gamma_{j,l}}{\delta^2} - N_s\right)\right\}$$

$$Set_J = \left\{j : j \in oldSet_J, \overline{\mathbf{Q}}_j(N_s) \geq \overline{\mathbf{Q}}_l(N_s) - W_{j,l}(N_s) \, \forall l \in oldSet_J, l \neq j\right\}$$

For every $j \in oldSet_J$ but $j \notin Set_J$, remove $(\mathbf{S}^n + \Delta_{i,j})$ from Set_C.

As shown in Algorithm 4.3, Step 0 initializes the parameters and variables used by the algorithm. Step 1 produces an estimate of the variance between the profits of two schedules and the upper bound statistics, N_j for $j \in Set_J$, to be used in subsequent steps to select the best schedule. There are two ways that a candidate schedule is selected (Step 2 and Step 3). One is when the sample size is sufficiently large to differentiate the best (or near-best) (Step 2). A schedule is considered near-

best, if its expected profit is within δ unit of the true best schedule. The value of δ is set by the decision-maker, and it is the the maximum difference between two profit values that are considered as equivalent in practice, namely, the indifference value. In this case, we simply select the schedule with the highest estimated profit. The second way is to screen out inferior schedules using a confidence interval, denoted by the statistic $W_{\cdot,\cdot}(\cdot)$ (Step 3), which is a function of the number of sampled batches. The indifference value is incorporated in $W_{\cdot,\cdot}(\cdot)$, and the larger the indifference value the smaller the $W_{\cdot,\cdot}(\cdot)$, leading to a more aggressive elimination process. Note that R&S holds for comparisons with analytically calculated data because they are treated as identical observations with zero variance. By KN, we are guaranteed with $(1-\alpha)$ percent confidence that the selected schedule for each patient request has an expected profit within δ unit of the best attainable schedule at that stage. We refer the readers to Kim and Nelson (2001) for the proof and recommendation on parameter selections (e.g., h^2).

4.5. Policy Comparison

In this section, we compare the RC policy with the Silo, Static and UC policies. Under a Silo policy, each clinic independently schedules patients to maximize its own objective. As mentioned in Section 4.1, referrals from APC to IMC are handled as unplanned add-ons in subsequent time slots in IMC's schedule with no advance coordination, resulting in more congestion at IMC. Despite the drawbacks of the Silo policy and the fact that healthcare services are highly interdependent, the Silo policy is common in practice. This is due to limiting methodology to coordinate

multi-clinic scheduling and clinics like APC and IMC are often separately managed services. To implement the Silo policy, we use the myopic approach of Muthuraman and Lawley (2008) at each clinic. Using a myopic approach for both Silo and coordinated policies allows us to better assess the impact of coordination on the objective function.

Under the Static policy both clinics still use the Silo policy, but prior to scheduling IMC blocks slots in the latter part of its schedule from exogenous IMC requests. The main reasons that the latter slots are blocked are: (1) to create a buffer in the IMC schedule that can benefit all APC referrals throughout the day and (2) to avoid clinical overtime. The number of slots blocked equals the estimated number of referrals to arrive during the day from APC, which is $J \cdot n_1 \cdot P_{1,2}$, rounded to the nearest integer. Different from RC which dynamically allocates patient assignments based on the current schedule, the location and the number of slots to be blocked are predetermined under a Static policy, even though referral patients arrive throughout the day. Due to its simplicity to implement, the Static policy of leaving predetermined "gaps in the schedule" to accommodate unplanned add-ons is also common in practice, even for separately managed services.

The UC policy does not restrict patients of a certain type. Without a booking limit, the schedules under UC can vary significantly based on different call-in sequences. Consequently, the scheduler risks making a local best decision which leads to a suboptimal final schedule.

Using numerical experiments, we compare RC with each of the alternative policies. We evaluate their performances based on the average, standard deviation and

coefficient of variance (CV) of the network profit. The variance of profit consists of the variation in the solution schedules across different call-in sequences, and the variation in a schedule's realized profit due to the uncertainties of patient no-shows, service times and referral flows. A favorable policy should exhibit both high profit and low variation,i.e., low CV.

We start with a base case of the PCSH where the parameter values are derived from the real system. There are two clinics, APC and IMC ($I = 2$) and 8 appointment slots each day ($J = 8$). APC has two residence providers that work in parallel and IMC has one provider (i.e., $n_1 = 2$, $n_2 = 1$). Their service times are exponentially distributed with identical service rate, which is normalized to 1. The attendance rates at both clinics are 80% ($\rho_1 = \rho_2 = 80\%$) and 25% of APC patients are referred to IMC ($P_{1,2} = 25\%$). The reward for service is set to a nominal value of 100 for both clinic ($r_1 = r_2 = 100$) and all costs are defined relative to these nominal values. Because of referrals, an APC patient brings an expected bundled reward of $r_1 + r_2 P_{1,2}$. To ensure patient satisfaction, the management of the PCSH would like their patients to wait no more than two slots. Therefore, we set the coefficient of overflow cost to be 0.5 of the service reward, i.e. $c_{i,j} = 0.5 r_i$, $j < J$. Patients that overflow from slot J incur clinical overtime, which is more expensive than the reward. We set the penalty for overtime as $c_{1,J} = 1.5(r_1 + r_2 P_{1,2})$ and $c_{2,J} = 1.5 r_2$. For the call-in process, we assume equal probability that a call requests APC (type 1 request) or IMC (type 2 request).

In addition to the base case, we examine the impact of differing referral rates, show-up rates and the call-in probabilities, on the solution of each policy. According

to the management of the PCSH, their current referral rate is 25% and they do not expect it to go above 55%. So we consider 25%, 35%, 45% and 55% for $P_{1,2}$. In terms of patient no-show, historical data reveals that it is between 10% to 25%. So we include show-up rates (90%, 90%), (80%, 80%) and (70%, 70%) in the study and consider their heterogenous combinations, i.e. (90%, 80%), (90%, 70%), (80%, 90%), (80%, 70%), (70%, 90%) and (70%, 80%). Lastly, we vary the probability of a type 1 request from 50%, to 25% and 75%.

For the simulation algorithm, we use 2000 randomly generated call-in sequences. The significant level and the indifference value are chosen at $\alpha = 0.01$ and $\delta = 2$ for R&S. The value of each schedule is simulated over 5,000 replications and we batch the data (500 per batch) to comply to the normality assumption in R&S, which passed the Shapiro-Wilk Normality Test with significance level 0.05. For each re-sampling, another 10 batches of data is generated. The choice of 10 batches per sampling is to optimize the computation time for R&S. Based on the choices of α, δ, and batch and sample sizes, all the simulation results in the following subsections are within 5% of the true best at a confidence level of at least 95%.

4.5.1 Silo

Under the Silo policy, the clinic only focuses on its own patient flows and overlooks the impact of referral patients. As a result, the overall network tends to be very congested. We evaluate the joint schedule of APC and IMC on the network objective function (Equation 4.4) to calculate the expected profit on the appointment day. Using the simulation procedure in Section 4.4, we obtain estimates of the average

profit, $\widetilde{V}$, the standard deviation of the profit, $\tilde{\sigma}$, and the coefficient of variation, $\widetilde{CV}$. These statistics versus referrals, show-up rates and call-in probabilities are shown in Figure 4.2, 4.3, and 4.4, respectively.

Consider the impact of referral rates on the RC results first. A higher referral rate indicates that APC receives more complicated patients that demand more resources in IMC. Therefore, fewer IMC and APC requests can be accommodated and the average profit decreases (Figure 4.2 (A)). Moreover, a higher referral rate increases the opportunity cost of each unfulfilled APC appointment (i.e. $r_1 + r_2 P_{1,2}$) due to no-show, and thus increases the uncertainty in the system (Figure 4.2 (B)). CV increases as the result of decreasing profit and increasing variance (Figure 4.2 (C)).

As for Silo, we observe a similar trend as in RC, with RC dominating the performance of Silo. The performance gap between these two policies enlarges from an improvement of 38.5% to 360% in average profit, and a reduction from 68.5% to 73% in profit uncertainty, as referral rate increases. These results shed light on the importance of coordinated scheduling. As our healthcare system becomes more patient-centric, different healthcare services become more interdependent as they endeavor to collaboratively deliver care in an integrated episode. As revealed in Figure 4.2, the operational inefficiency is too costly for practitioners to continue with a Silo policy.

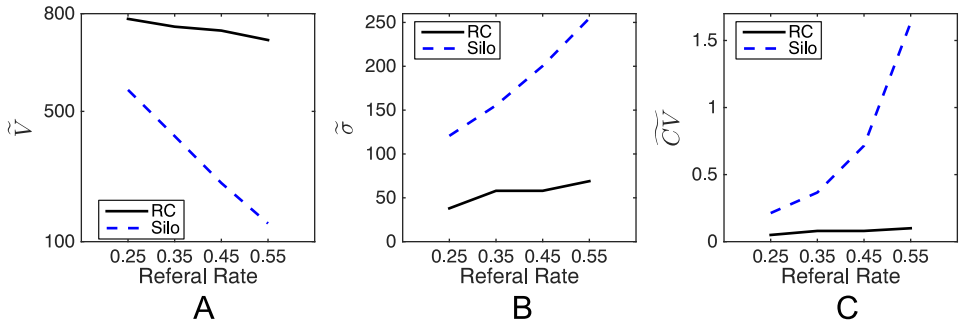

Figure 4.2: RC v.s. Silo: Referral Rate

In terms of patient no-shows, as the no-show rate increases (ρ_1 or ρ_2 decreases), the profit under RC decreases (Figure 4.3 (A)) while its standard deviation and CV increase (Figure 4.3 (B) and (C), respectively). These results reflect the detrimental impact of patient no-shows. A higher no-show rate introduces more uncertainty into the PCSH and thus increases the system variation and decreases the profit. Moreover, the show-up rate at APC (ρ_1) has more impact on the profit than that of IMC (ρ_2). For example, an increase of ρ_1 from 70% to 90% increases the profit by 6%, when $\rho_2 = 70\%$. While, the profit only increases by 2% when ρ_2 increases from 70% to 90%, at $\rho_1 = 70\%$. In addition, the average profit at show-up rates $(90\%, 70\%)$ is higher than that at (70%, 90%) and (80%, 90%). This result indicates that patient no-shows at clinics with more complex patients (i.e., APC) are more operational costly and hence should be managed more carefully.

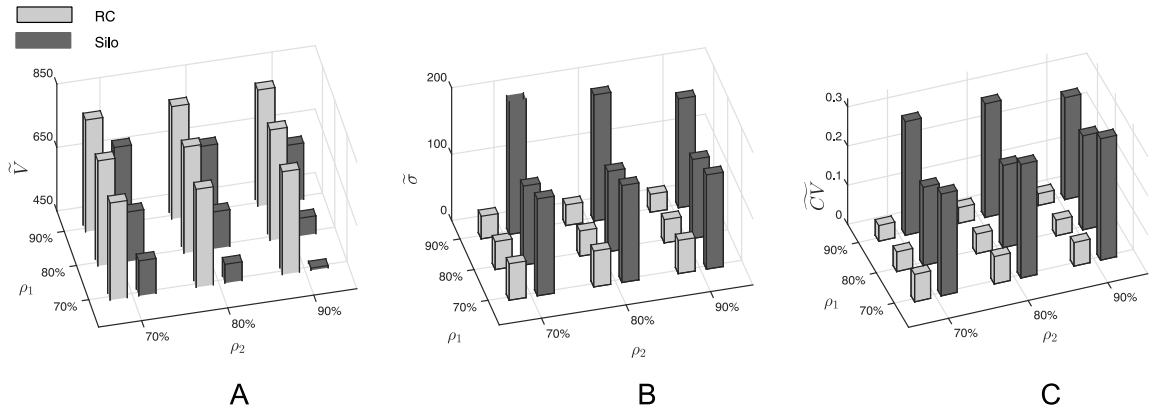

Figure 4.3: RC v.s. Silo: Show-up Rates

Similar to RC, the average profit under Silo also decreases as ρ_1 decreases, for ρ_2 fixed. However, in contrast to RC, the average profit under Silo increases with decreasing ρ_2, when ρ_1 is fixed. The reason is that under Silo, both APC and IMC are scheduled to full capacity, which results in an overly congested network, especially at IMC. Hence a huge operational cost is incurred on the appointment day. But with a higher no-show rate at IMC, the congestion of the network is moderately alleviated if fewer patients show up. As shown in Figure 4.3 (A), the improvement from Silo to RC ranges from 14% to 35% at $\rho_2 = 70\%$, but the range increases to 30% to 70% at $\rho_2 = 90\%$. In term of the profit variation, by changing from Silo to RC, the reduction in system risks is quite substantial, ranging from 65% to 83%.

Lastly, we examine how the probability of type 1 requests affect the solution. A higher rate of type 1 requests has no effect on Silo, but slightly increases the profit under RC. For the Silo policy, its schedule is independent of the call-in sequences so there is a single solution, represented by a flat dashed line in Figure 4.4. As for RC,

because of the booking limit, the dependency of RC on different call-in sequences is mitigated. Even still, when the rate of type 1 request is higher, the scheduler is more likely to receive APC requests early to better balance the utilization of the PCSH and to achieve a higher system profit. Overall, RC is robust and outperforms Silo with significantly higher system profit and lower variation.

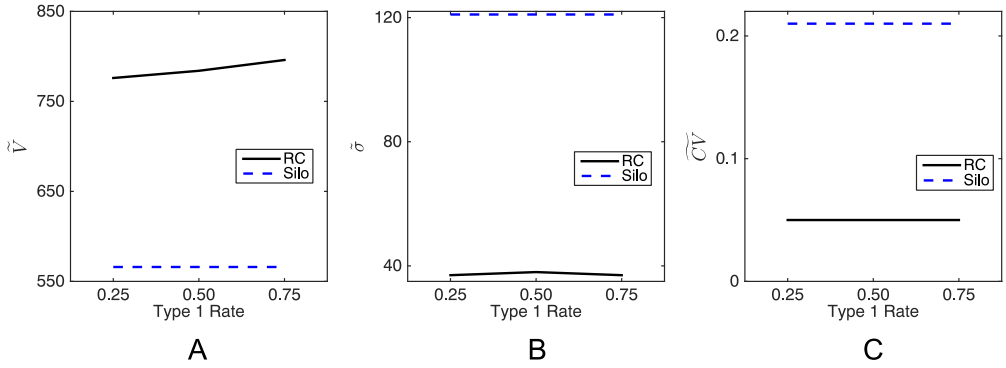

Figure 4.4: RC v.s. Silo: Call-in Sequence

In summary, the comparison of RC with Silo in the PCSH addresses the first set of research questions posed in the introduction of this chapter. Silo does not appear to be a sustainable policy across a wide range of parameter values that might be expected in practice. Additionally, we have demonstrated that the opportunity cost of Silo versus RC can be substantial in terms of average performance and risk, as measured by the standard deviation of performance.

4.5.2 Static

Recall that, under Static, APC and IMC schedule patients independently using the Silo policy. In anticipation of the increasing workload due to referral patients

on the appointment day, IMC blocks slots in the latter part of its schedule for APC referrals and only schedules its patients in the early slots. Such an approach is a single-sided attempt by IMC to accommodate referrals and minimal collaboration with APC is needed. However, as referral patients can arrive at IMC throughout the day, packing IMC patients in the early slots may cause an unbalanced utilization of IMC resources at different times of the day. As shown in Figure 4.5 to 4.7, the Static policy is less efficient in balancing the workload of the PCSH, compared to RC.

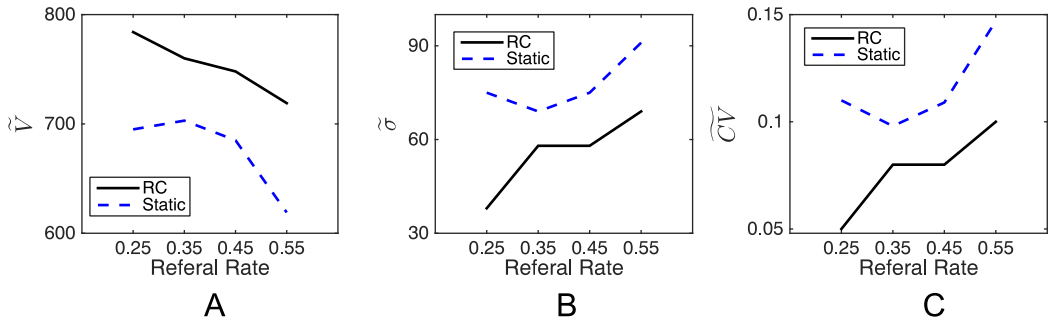

Figure 4.5: RC v.s. Static: Referral Rate

As the referral rate increases, the performance of Static first improves (a slight increase in profit and a slight decrease in standard deviation and CV), then degrades (above 35% referral rate). The degrade in performance is driven by two factors: (1) Silo overpopulates APC and (2) as the referral rate increases, IMC runs out of the slots to be blocked for APC referrals. In comparison, the advantages of RC are 16%, 24% and 35% over Static, in terms of the average, standard deviation and CV of the profit at 55% referral rate, respectively.

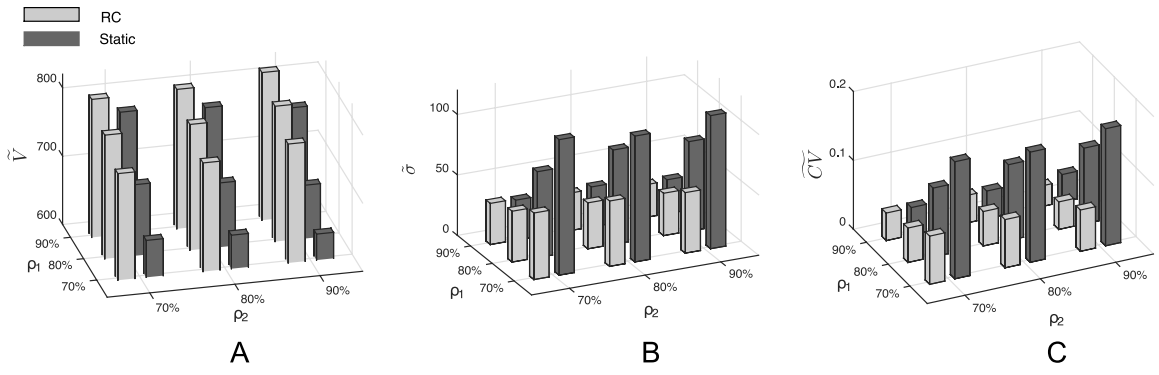

Figure 4.6: RC v.s. Static: Show-up Rates

In terms of patient no-shows, we expect a similar pattern as in Silo, because Static is a Silo policy applied to fewer appointment slots at IMC. Although, IMC limits the number of the exogenous IMC patients in its schedule, the overly congested APC still leads to a crowded IMC on the appointment day and thus, burdens the PCSH with high cost penalties and system risks. Lastly, as shown in Figure 4.7, the Static policy is independent of the call-in process, and is significantly outperformed by RC.

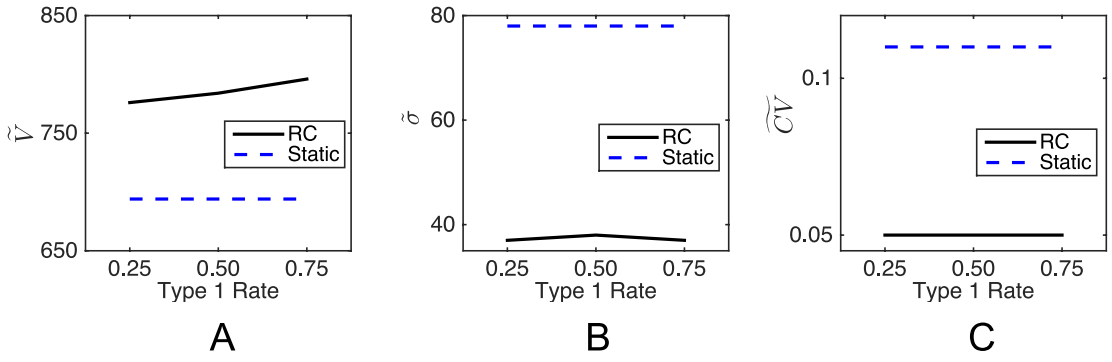

Figure 4.7: RC v.s. Static: Call-in Sequence

To sum up, in this subsection, we answer the second research question on the effectiveness of the Static policy. Static improves upon Silo. However, its effectiveness in coordinating patient scheduling is rather limited compared to RC, which outperforms Static over all parameter values tested in this study. Moreover, the advantage of RC over Static is increasingly large as referral probability increases from medium to high (Figure 4.5 (A) to (C)), or as ρ_1 decreases (Figure 4.6 (A) to (C)). By adopting RC rather than Static, we expect improvements in the operational efficiency and more robust performance for the PCSH, especially when the referral rate is high and the attendance rate at APC is low.

4.5.3 UC

By comparing RC and UC, we show that a simple booking limit on the IMC requests can increase the performance and the robustness of a coordinated myopic policy. Looking across Figures 4.8, 4.9 and 4.10, we observe that the results under UC exhibit similar trends as in RC, with RC dominating UC across all parameter

values and on all three performance metrics.

Firstly, a higher referral rate decreases the average profit but increases the variation and CV of the profit for both RC and UC. However, the performance of UC degrades much faster than RC. The reason is that UC, without limiting the number of IMC requests, is more likely to construct a suboptimal schedule, especially when the referral rate is high and the early requests are dominated by type 2. On the contrary, RC avoids overpopulating IMC under such sequences and its results are hence more robust.

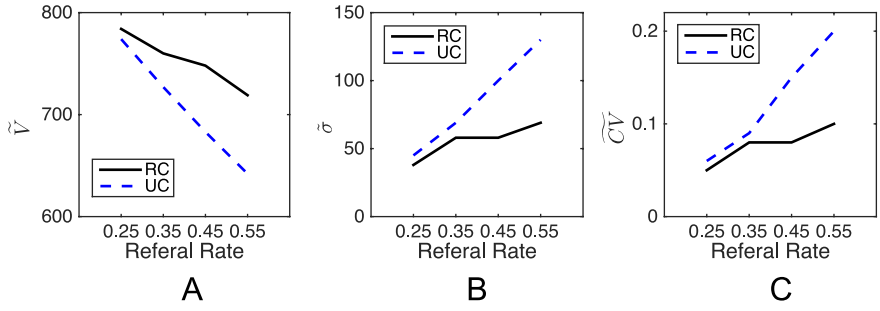

Figure 4.8: RC v.s. UC: Referral Rate

In terms of no-shows, the advantage of RC over UC is more obvious at higher attendance rates (Figure 4.9). For example, the reduction in the profit standard deviation, when switching from UC to RC, is over 25% at (90%,90%) show-up rates, but is only 5% at (70%,70%). An insight from this result is that to gain more relative benefit from RC, the management needs to focus on increasing the patients' attendance rate.

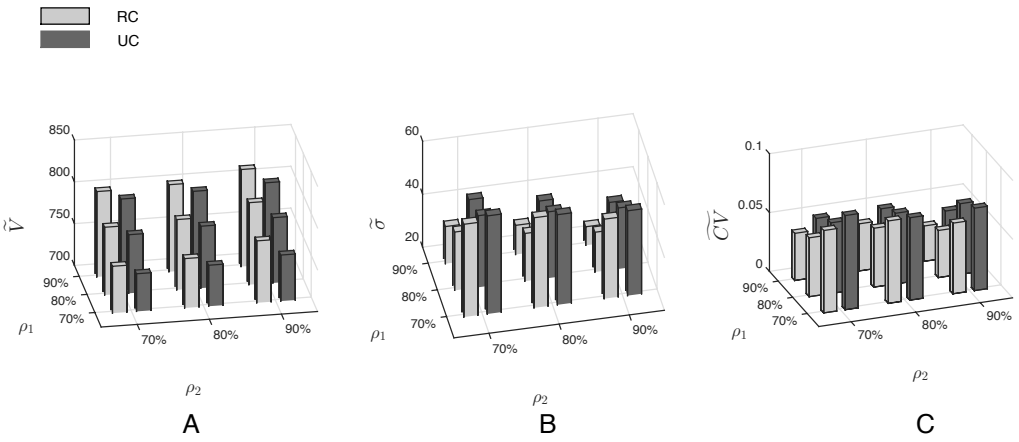

Figure 4.9: RC v.s. UC: Show-up Rates

Lastly, the results in Figure 4.10 confirms the necessity to implement the booking limit. Across all type 1 rates, RC has a much robust solution than UC, in the average, standard deviation and CV of the profit. For example, at 25% type 1 rate, RC leads to a reduction in system variation over 20%, compared to UC.

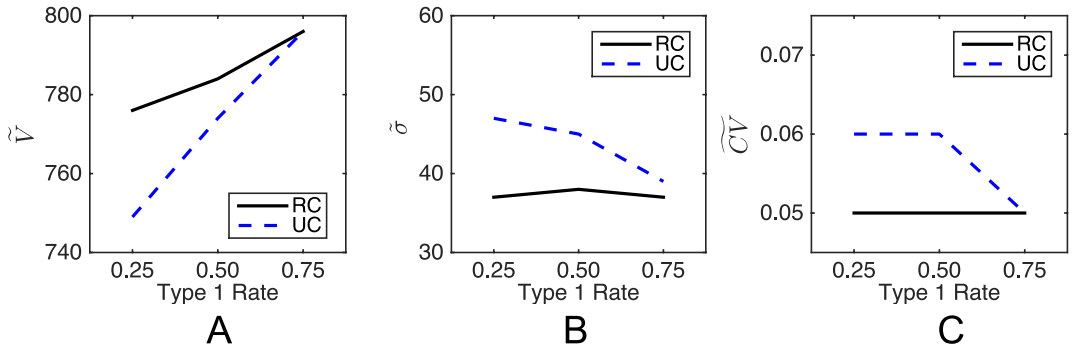

Figure 4.10: RC v.s. UC: Call-in Sequence

4.6. Concluding Remarks

In this chapter, we formulate a multi-server, multi-clinic PCSH model to coordinate different outpatient services, where multiple providers work in parallel in each service. Our goal is to reduce the operational inefficiencies and system risks in the PCSH. We develop a joint-scheduling method and propose a coordinated scheduling policy with booking limit to sequentially allocate patient appointment requests. In a coordinated PCSH that guarantees same day referrals, our approach yields high quality, robust solutions.

By comparing our fully coordinated policy against other policies being considered by the PCSH, we answer the questions posed in Section 4.1. First, the PCSH with same-day referrals is not sustainable under a Silo policy, where the system would become highly congested and inefficient. Second, although the Static policy improves the Silo policy, it cannot coordinate APC and IMC as effectively and efficiently as the fully coordinated RC policy. Third, the RC policy results in high profit, low risk, robust solutions that outperform Silo, Static and UC across a wide range of parameter values examined in this study.

As our healthcare system transforms to value-based, patient-centric care, the interdependencies among different services to collaboratively deliver an episode of care increase. Therefore, the potential impact of our coordinated scheduling method can be great. Moreover, our scheduling algorithm and simulation procedures provide a framework that can be extended to other collaborative networks with different modeling assumptions, such as the service time distribution and the arrival time of referrals.

In future research, we can relax the service time distribution from exponential to lognormal, which is believed by many researchers to be more realistic for certain healthcare services (e.g., Cayirli et al. (2006), Zacharias and Pinedo (2014)), and other distributions. The scheduling algorithm is independent of the service distribution, however, without the memerylessness of the exponential distribution, we lose analytical tractability. As a result, we would have to resort to an all-simulation approach, without the hybrid evaluation method to expedite the scheduling algorithm. Similarly, the assumption that referral patients arrive at the immediate subsequent slot at IMC, provides analytical tractability to our model. In future studies, we could use simulation to evaluate the impact of deferring referral services at later slots in IMC (e.g., all referral patients arrive at the last slot of IMC), and compare different referral rules.

Appendices

Appendix A

Evaluating the effectiveness of government subsidy on the adoption of energy efficient durable products

A.1. Contrast Between Two Proportions Based on Individually Paired Data

Using the notation of Newcombe(1998), let n represent the total number of patients in the study, e be the number of ASA 1 and 2 patients not sent to APC, f be the number of ASA 3 and 4 patients not sent to APC, g be the number of ASA 1 and 2 patients sent to APC, and h be the number ASA 3 and 4 patients sent to APC. We wish to calculate a confidence interval for a quantity θ which is the difference between the proportion of ASA 3 and 4 patients not sent to APC and the proportion of ASA 1 and 2 patients sent to APC. Using the large sample approximation, an approximate 95% confidence interval for θ is $\frac{f-g}{n} \pm z_{0.975}\sqrt{\frac{(e+h)(f+g)+4fg}{n^3}}$ where $z_{0.975}$ is the 0.975 fractile of the standard Normal distribution.

A.2. Randomized Test Likelihood Calculations

The total number of ways in which 7 delays can happen with 79 patients is $\begin{pmatrix} 79 \\ 7 \end{pmatrix}$. Segmenting the data set into APC and Non-APC patients, the number of ways in which six or more of the seven delays happen for Non-APC patients is

$\binom{38}{6}\binom{41}{1}+\binom{38}{7}\binom{41}{0}$. Hence, the probability of six or more of the seven delays happening to Non-APC patients randomly is $\frac{\binom{38}{6}\binom{41}{1}+\binom{38}{7}\binom{41}{0}}{\binom{79}{7}} =$ 0.0434.

Bibliography

Ahmadi-Javid, A., Z. Jalali, and K. Klassen (2016). Outpatient appointment systems in healthcare: A review of optimization studies. *European Journal of Operational Research*.

Alexopoulos, C., D. Goldsman, J. Fontanesi, D. Kopald, and J. R. Wilson (2008). Modeling patient arrivals in community clinics. *Omega 36*, 33–43.

Arena (2013). Arena simulation software. `http://www.areansimulation.com/Arena-Home.aspx/`. Last accessed: 2013-05-21.

ASA (2011). The perioperiative or surgical home. Proposal to the ASA House of Delegates 2011 Session, August 21, 2011.

ASA (2013). The American Society of Anesthesiologists. `http://www.asahq.org/Home/ForMembers/Clinical-Information/ASAPhysicalStatusClassificationSystem/`. Last accessed: 2013-05-21.

Bailey, N. (1952). A study of queues and appointment systems in hospital outpatient departments, with special reference to waiting times. *Journal of the Royal Statistical Society - Series B 14*(2), 185–199.

Begen, M. A. and M. Queyranne (2011). Appointment scheduling with discrete random durations. *Mathematics of Operations Research 36*, 240–257.

Berg, B. and B. T. Denton (2012). Appointment planning and scheduling in outpatient procedure centers. In *Handbook of Healthcare System Scheduling*, pp. 131–154. Springer, U.S.

Bureau of Economic Analysis (2016). National income and product accounts: gross domestic product, fourth quarter and annual 2015 (advance estimate). http://www.bea.gov/newsreleases/national/gdp/gdpnewsrelease.htm/. Last accessed: 2016-02-06.

Cayirli, T. and E. Veral (2003). Outpatient scheduling in health care: a review of literature. *Production and Operations Management 12*(4), 519–549.

Cayirli, T., E. Veral, and H. Rosen (2006). Designing appointment scheduling systems for ambulatory care services. *Health Care Management Science 9*(1), 47–58.

Cayirli, T., K. K. Yang, and S. A. Quek (2012). A universal appointment rule in the presence of no-shows and walk-ins. *Production and Operations Management 21*(4), 682–697.

Centers for Medicare and Medicaid Services (2009). Roadmap for Implementing Value Driven Healthcare in the Traditional Medicare Fee-for-Service Program. http://www.cms.gov/Medicare/QualityInitiativesPatientAssessment Instruments/QualityInitiativesGenInfo/downloads/vbproadmap-oea-1-16-508.pdf/. Last accessed: 2016-04-04.

Centers for Medicare and Medicaid Services (2014). National health expenditure data projections 2014-2024 - forecast summary. http://www.cms.gov/

ResearchStatisticsDataAndSystems/StatisticsTrendsAndReports/NationalHealthExpendData/NationalHealthAccountsProjected.html/. Last accessed: 2016-02-06.

Centers for Medicare and Medicaid Services (2015). National Health Expenditure Projections, 2014-24: Spending Growth Faster Than Recent Trends. http://www.ncbi.nlm.nih.gov/pubmed/26220668/. Last accessed: 2016-04-04.

Chakraborty, S., K. Muthuraman, and M. Lawley (2010). Sequential clinical scheduling with patient no-shows and general service time distributions. *IIE Transactions 42*(5), 354–366.

Chao, X., L. Liu, and S. Zheng (2003). Resource allocation in multisite service systems with intersite customer flows. *Management Science 49*(12), 1739–1752.

Chen, R. R. and L. W. Robinson (2014). Sequencing and scheduling appointments with potential call-in patients. *Production and Operations Management 23*(9), 1522–1538.

Chen, Z. and N. G. Hall (2007). Supply chain scheduling: conflict and cooperation in assembly systems. *Operations Research 55*(6), 1072–1089.

Correll, D., A. Bader, M. Hull, C. Hsu, L. Tsen, and D. Hepner (2006). Value of preoperative clinic visits in identifying issues with potential impact on operating room efficiency. *Anesthesiology 105*, 1254–1259.

Cullen, K., M. Hall, and A. Golosinsky (2009). Ambulatory surgery in the united states, 2006. *National Health Stat R* (11).

Dawande, M., H. N. Geismar, N. G. Hall, and C. Sriskandarajah (2009). Supply chain scheduling: distribution systems. *Production and Operations Management 15*(2), 243–261.

Denton, B. and D. Gupta (2003). A sequential bounding approach for optimal appointment scheduling. *IIE Transactions 35*(11), 1003–1016.

Dexter, F. (1999). Design of appointment systems for preanesthesia evaluation clinics to minimize patient waiting times: A review of computer simulation and patient survey studies. *Anesthesia and Analgesia 89*, 925–931.

Eagle, K., P. Berger, and H. Clakins (2002). Acc/aha guideline update for perioperative cardiovascular evaluation for noncardiac surgery â executive summary. *Anesthesia and Analgesia 94*(5), 1052–1064.

EasyFit (2013). Mathwave: Data analysis and simulation software. http://www.mathwave.com /products/easyfit.html. Last accessed: 2013-05-21.

Edward, G., S. Das, S. Elkhuizen, P. Bakker, J. Hontelez, M. Hollmann, B. Precekl, and L. Lemaire (2008). Simulation to analyse planning difficulties at the preoperative assessment clinic. *British Journal of Anaesthesia 100*(2), 195–202.

Erdogan, S. A. and B. Denton (2013). Dynamic appointment scheduling of a stochastic server with uncertain demand. *INFORMS Journal on Computing 25*(1), 116–132.

Erdogan, S. A., A. Gose, and B. Denton (2015). Online appointment sequencing and scheduling. *IIE Transactions 47*(11), 1267–1286.

Feldman, J., N. Liu, H. Topaloglu, and S. Ziya (2014). Appointment scheduling under patient preference and no-show behavior. *Operations Research 64*(4), 794–811.

Ferschl, M., A. Tung, B. Sweitzer, D. Huo, and D. Glick (2005). Preoperative clinic visits reduce operating room cancellations and delays. *The Journal of the American Society of Anesthesiologists 103*(4), 855–859.

Gibby, G. and W. Schwab (1998). Availability of records in an outpatient preanesthetic evaluation clinic. *Journal of Clinical Monitoring and Computing 14*, 385–391.

Glowacka, K., R. Henry, and J. May (2009). A hybrid data mining/simulation approach for modelling outpatient no-shows in clinic scheduling. *Journal of the Operational Research Society 60*(8), 1056–1068.

Gupta, D. and B. Denton (2008). Appointment scheduling in health care: challenges and opportunities. *IIE Transactions 40*(9), 800–819.

Gupta, D. and W. Wang (2012). Patient appointments in ambulatory care. In *Handbook of Healthcare System Scheduling*, pp. 65–104. Springer, U.S.

Halaszynski, T., R. Juda, and D. Silverman (2004). Optimizing postoperative outcomes with efficient preoperative assessment and management. *Critical Care Medicine 32*(4), S76–S86.

Hall, N. G. and C. N. Potts (2003). Supply chain scheduling: batching and delivery. *Operations Research 51*(4), 566–584.

Hassin, R. and S. Mendel (2008). Scheduling arrivals to queues: a single-server model with no-shows. *Management Science 54*(3), 565–572.

Health Cost Containment and Efficiencies (2010). Episode-of-care payments. Technical report, National Conference of State Legislatures.

Hepner, D., A. Bader, S. Hurwitz, M. Gustafson, and L. Tsen (2004). Patient satisfaction with preoperative assessment in a preoperative assessment testing clinic. *Anesthsia and Analgesia 98*, 1099–1105.

Judge, G., R. Hill, W. Griffiths, H. Lutkepohl, and T.-S. Lee (1988). *Introduction to the Theory and Practice of Econometrics, 2nd ed.* John Wiley & Sons.

Kaandorp, G. C. and G. Koole (2007). Optimal outpatient appointment scheduling. *Health Care Management Science 10*(3), 217–229.

Kelton, W., R. Sadowski, and N. Swets (2010). *Simulation with Arena, 5th ed.* McGraw-Hill.

Kim, S. and B. Nelson (2001). A fully sequential procedure for indifference-zone selection in simulation. *ACM Transactions on Modeling and Computer Simulation (TOMACS) 11*(3), 251–273.

Klassen, K. J. and T. R. Rohleder (1996). scheduling outpatient appointments in a dynamic environment. *Journal of Operations Management 14*(2), 83–101.

Klassen, K. J. and R. Yoogalingam (2009). Improving performance in outpatient appointment services with a simulation optimization approach. *Production and Operations Management 18*(4), 447–458.

LaGanga, L. R. and S. R. Lawrence (2007). Clinic overbooking to improve patient access and increase provider productivity. *Decision Sciences 38*(2), 251–276.

LaGanga, L. R. and S. R. Lawrence (2012). Appointment overbooking in health care clinics to improve patient service and clinic performance. *Production and Operations Management 21*(5), 874–888.

Lahiri, A. and A. Seidmann (2012). Information hang-overs in healthcare service systems. *Manufacturing and Service Operations Management 14*(4), 634–653.

Law, A. and W. Kelton (2010). *Simulation Modeling and Analysis, 3rd ed.* McGraw-Hill.

Lin, J., K. Muthuraman, and M. Lawley (2011). Optimal and approximate algorithms for sequential clinical scheduling with no-shows. *IIE Transactions on Healthcare Systems Engineering 1*(1), 20–36.

Liu, L. and X. Liu (1998). Block appointment systems for outpatient clinics with multiple doctors. *Journal of the Operational Research Society 49*(12), 1254–1259.

Lloyd, C. (1990). Confidence intervals from the difference between two correlated proportions. *Journal of the American Statistical Association 85*(412), 1154–1158.

Luo, J., V. G. Kulkarni, and S. Ziya (2012). Appointment scheduling under patient no-shows and service interruptions. *Manufacturing and Service Operations Management 14*(4), 670–684.

Merritt Hawkins and Associates (2009). 2009 Survey of Physician Appointment Wait Times. http://www.merritthawkins.com/pdf/mha2009waittimesurvey.pdf/. Last accessed: 2016-04-04.

Millhiser, W., E. Veral, and B. Valenti (2012). Assessing appointment systemsâ operational performance with policy targets. *IIE Transac-tions on Healthcare Systems Engineering 2*, 274–289.

Morrice, D., D. Wang, J. Bard, L. Leykum, S. Norrily, and P. Veerapaneni (2013). A simulation analysis of a patient-centered surgical home to improve outpatient surgical pro-cesses of care and outcomese. *Proceedings of the 2013 Winter Simulation Conference*, 2274–2286.

Morrice, D. J., D. Wang, J. F. Bard, L. K. Leykum, S. Noorily, and P. Veerapaneni (2014). A patient-centered surgical home to improve outpatient surgical processes of care and outcomes. *IIE Transactions on Healthcare Systems Engineering 4*, 119–134.

Muthuraman, K. and M. Lawley (2008). A stochastic overbooking model for outpatient clinical scheduling with no-shows. *IIE Transactions 40*(9), 820–837.

Newcombe, R. (1998). Improved confidence intervals for the difference between binomial proportions based on paired data. *Statistics in Medicine 17*, 2635–2650.

Newman, M., J. Mathew, and S. Aronson (2013). The evolution of anesthesiology and perioperative medicine. *Anesthesiology 118*(5), 1005–1007.

Palisade (2013). Palisade stattools: Advance statistical analysis for excel. http://www.palisade.com/stattools/. Last accessed: 2013-05-21.

Parker, G. and E. Anderson Jr. (2002). From buyer to integrator: Thetransformation of the supply chain manager in the vertically disintegrating firm. *Production and Operations Management 11*(1), 75–91.

Porter, M. (2009). A strategy for health care reformâtoward a value-based system. *New England Journal of Medicine 361*(2), 109–112.

Robinson, L. W. and R. R. Chen (2010). A comparison of traditional and open-access policies for appointment scheduling. *Manufacturing and Service Operations Management 12*(2), 330–346.

Rohleder, R., P. Lewkonia, D. Bischak, P. Duffy, and R. Hendijani (2011). Using simulation modeling to improve patient flow at an outpatient orthopedic clinic. *Health Care Management Science 14*, 135–145.

Sickinger, S. and R. Kolisch (2009). The performance of a generalized bailey–welch rule for outpatient appointment scheduling under inpatient and emergency demand. *Health care management science 12*(4), 408–419.

Stange, K. C., P. Nutting, W. L. Miller, C. R. Jaén, B. F. Crabtree, S. Flocke, and J. M. Gill (2010). Defining and measuring the patient-centered medical home. *Journal of General Internal Medicine 25*(6), 601–12.

STATA (2013). Stata: Data analysis and statistical software. http://www.stata.com/. Last accessed: 2013-05-21.

Tsen, L., S. Segal, M. Pothier, L. Hartley, and A. Bader (2002). The effect of alterations in a preoperative assessment clinic on reducing the number and improving the yield of cardiology consultations. *Anesthsia and Analgesia 95*, 1563–1568.

VA Report (2012). Access and coordination of care at harlingen community based outpatient clinic VA Texas valley coastal bend healthcare system Harlingen, Texas. Technical report, Office of Inspector General, Veterans Health Administration.

VA Report (2014). Review of patient wait times, scheduling practices, and alleged patient deaths at the Phoenix healthcare system. Technical report, Office of Inspector General, Veterans Health Administration.

van Klei, W., K. Moons, C. Rutten, A. Schuurhuis, J. Knape, C. Kalkman, and D. Grobbee (2002). The effect of outpatient preoperative evaluation of hospital inpatients on cancellation of sur-gery and length of hospital stay. *Anesthsia and Analgesia 94*, 644–649.

Vetter, T., L. Goeddel, A. Boudreaux, T. Hunt III, K. Johnes, and J.-F. Pittet (2013). The perioperative surgical home: how can it make the case so everyone wins? *BMC Anesthesiology 13*(6).

Vetter, T., N. Ivankova, L. Goeddel, G. McGwin, and J.-F. Pittel (2013). An analysis of methodologies that can be used to validate if a perioperative surgical home improves the patient-centeredness, evidence-based practice, quality, safety, and value of pa-tient care. *Anesthesiology 119*(6), 1261–1274.

Vetter, T. R., A. M. Boudreaux, K. A. Jones, J. M. Hunter, and J. Pittet (2014). The perioperative surgical home: how anesthesiology can collaboratively achieve and leverage the triple aim in health care. *Anesthesia and Analgesia 118*(5), 1131–6.

Wang, P. P. (1993). Static and dynamic scheduling of customer arrivals to a single-server system. *Naval Research Logistics 40*(3), 345–360.

White, D. L., C. M. Froehle, and K. J. Klassen (2011). The effect of integrated scheduling and capacity policies on clinical efficiency. *Production and Operations Management 20*(3), 442–455.

Zacharias, C. and M. Pinedo (2014). Appointment scheduling with no-shows and overbooking. *Production and Operations Management 23*(5), 788–801.

Zacharias, C. and M. Pinedo (2016). Managing customer arrivals in service systems with multiple servers. Working paper.

Zeng, B., A. Turkcan, J. Lin, and M. Lawley (2009). Clinic scheduling models with overbooking for patients with heterogeneous no-show probabilities. *Annals of Operations Research 178*(1), 121–144.

Zonderland, M., F. Boer, R. Boucherie, and J. v. K. de Roode (2009). Redesign of a university hospital preanesthesia evaluation clinic using a queuing theory approach. *Anesthsia and Analgesia 109*(5), 1612–1621.